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[7,8], 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

23 [8]Museum of Southwestern Biology and Department of Biology, University of New Mexico,

24 Albuquerque, NM 87131, USA

25

26 [*] Correspondence to be sent to: Diego F. Morales-Briones. Department of Plant and Microbial

27 Biology, University of Minnesota, 1445 Gortner Avenue, St. Paul, MN 55108, USA, Email:

28 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, especially in groups experiencing whole-genome duplication (WGD) events. Common approaches for processing paralogs in target enrichment datasets include random selection, manual pruning, and mainly, the removal of entire genes that show any evidence of paralogy. These approaches are prone to errors in orthology inference or removing large numbers of genes. By removing entire genes, valuable information that could be used to detect and place WGD events is discarded. Here we use an automated approach for orthology inference 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 ancient WGDs in the group. However, both the phylogenetic location and putative parental lineages of these WGD events 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 autopolyploid origin of the entire *Alchemilla* s.l. and the nested allopolyploid origin of four major clades within the group. Here we showed the utility of automated tree-based orthology inference methods, previously designed for 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.

52 Polyploidy, or whole genome duplication (WGD), is prevalent throughout the evolutionary
53 history of plants (Cui et al. 2006; Jiao et al. 2011; Jiao et al. 2012; Leebens-Mack et al. 2019). As
54 a result, plant genomes often contain large numbers of paralogous genes from recurrent gene and
55 genome duplication events (Lynch and Conery 2000; Panchy et al. 2016). Paralogs are defined as
56 homologous genes that share a common ancestor as the product of gene duplication (Fitch 1970),
57 either from small scale duplications or WGD. One special case of WGD is allopolyploidy, where
58 genome doubling is accompanied by hybridization between two different species. The duplicated
59 genes in allopolyploids are not paralogs in the traditional sense and are referred to as
60 homoeologs, which are expected to be sister to the orthologous in the parental taxa, rather than to
61 each other (Smedmark et al., 2003). For practical purposes, however, we refer to the product of
62 any kind of duplications found in gene trees hereafter as paralog, as homoeologs are
63 indistinguishable from paralogs until diagnosed as resulting from allopolyploidy. With very few
64 nuclear genes being truly single- or low-copy, careful evaluation of orthology is critical for
65 phylogenetic analyses (Fitch 1970). Orthology inference has received much attention in the
66 phylogenomic era with multiple pipelines available for this task (e.g., Li et al. 2003; Dunn et al.
67 2013; Kocot et al. 2013; Yang and Smith 2014; Emms and Kelly 2019, also see Glover et. al
68 2019 and Fernández et al. 2020 for recent reviews). But these approaches have been mainly
69 applied to genomic or transcriptomic data sets. So far, few studies have employed automated,
70 phylogeny-aware orthology inference in target enrichment datasets. The most common approach
71 for dealing with paralogy in target enrichment datasets is removing entire genes that show any
72 evidence of potential paralogy (e.g., Nicholls et al. 2015; Jones et al. 2019; Andermann et al.
73 2020; but see Moore et al. 2018). Removal of entire genes might seem appropriate in target
74 enrichment datasets in which only a small number of genes show evidence of paralogy (e.g.,

75 Larridon et al. 2020), but in some datasets this could result in a significant reduction of the
76 number of loci (e.g., Montes et al. 2019). More importantly, dealing with paralogy only by
77 removal of entire genes assumes that target enrichment assembly pipelines (e.g., Faircloth 2016;
78 Johnson et al. 2016; Andermann et al. 2018), have flagged all genes with paralogs. It also
79 assumes that if no sequence in a gene is flagged, all sequences are all single-copy and
80 orthologous. On the other hand, this approach also removes genes that show allelic variation
81 instead of paralogs. Given the prevalence of WGD and reticulations these assumptions can lead
82 to errors in orthology inference. As paralogous genes are prevalent in plants, more appropriate
83 orthology inference methods need to be applied in target enrichment data. The same automated
84 approaches used for genome and transcriptome datasets can be applied for target enrichment, as
85 these are tree-based and agnostic to the data source for tree inference.

86 The ability to explicitly process paralogs opens the door for using target enrichment data
87 for inferring gene duplication events and pinpointing the phylogenetic locations of putative
88 WGDs. In the past, the phylogenetic placement of WGD events have most often been carried out
89 using genome and transcriptome sequencing data (e.g., Li et al. 2015; Huang et al. 2016; McKain
90 et al. 2016; Yang et al. 2018) using either the synonymous distance between paralog gene pairs
91 (Ks; Lynch and Conery 2000) or tree-based reconciliation methods (e.g., Jiao et al. 2011; Li et
92 al. 2015; Yang et al. 2015; Huang et al. 2016; Xiang et al. 2017; Leebens-Mack et al. 2019).
93 Similar to orthology inference, tree-based methods used to investigate WGDs in genome and
94 transcriptome datasets should be useful in target enrichment data. Target enrichment methods
95 (e.g., Mandel et al. 2014; Weitemier et al. 2014; Buddenhagen et al. 2016) have been widely
96 adopted to collect hundreds to over a thousand nuclear loci for plant systematics, allowing
97 studies at different evolutionary scales (e.g., Villaverde et al. 2018), and the use of museum-

preserved collections (e.g., Forrest et al. 2019). This creates new opportunities to adopt tree-based reconciliation methods to explore WGD patterns in groups for which genomic and transcriptomic resources are not available or feasible.

With at least 350 (–1,100) species worldwide, *Alchemilla* in the broad sense has been a challenging group to study due to the presence of reticulate evolution, polyploidy, and apomixis . Based on previous phylogenetic analyses, *Alchemilla* s.l. contains four clades: Afromilla, *Aphanes,* Eualchemilla, and *Lachemilla* (Table S1). Together they form a well-supported clade nested in the subtribe Fragariinae, which also includes the cultivated strawberries (Gehrke et al. 2008). Unlike the more commonly recognized members of the rose family (Rosaceae), *Alchemilla* s.l. is characterized by small flowers with no petals, and a reduced number (1–4[–5]) of stamens that have anthers with one elliptic theca on the ventral side of the connective that opens by one transverse split (Perry 1929; Soják 2008). Gehrke et al. (2008) presented the first phylogeny of *Alchemilla* s.l. and established the paraphyly of traditional *Alchemilla* s.s. as consisted of a primarily African clade, Afromilla, and a Eurasian clade, Eualchemilla. Gehrke et al. (2008) also 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 S1). 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 in other genera of Fragariinae (Lundberg 2009; Eriksson et al. 2015; Gehrke et al. 2016, Kamneva et al. 2017; Morales-Briones

121 et al. 2018a), leaving the phylogenetic relationships of *Alchemilla* s.l. to the rest of Fragariinae
122 unresolved. Unlike most members of Fragariinae that have predominantly diploid species,
123 *Alchemilla* s.l. is known for high rates of polyploidy. The base chromosome number of
124 *Alchemilla* s.l. is eight (x = 8), which differs from all other members in Fragariinae that have a
125 base of number of seven (x = 7; Dickinson et al. 2007; Lundberg et al., 2009). Ploidy levels have
126 been well documented in Eualchemilla that shows only polyploid species ($2n$ = 64 to 220–224;
127 octoploid to 28-ploid; e.g., Turesson 1943; Izmailow 1981; Walters and Bozman, 1967;
128 Hayirhoğlu-Ayaz et al. 2006). *Aphanes* has mainly diploid species ($2n$ = 16), with the exception
129 of *Aphanes arvensis* that is an hexaploid ($2n$ = 48; Montgomery et al. 1997). *Lachemilla* has
130 mostly polyploid members ($2n$ = 24 to 96; triploid to 12-ploid) with a single species reported to
131 have diploid ($2n$ = 16) and triploid ($2n$ = 24) populations (Morales-Briones et al. 2018a). Lastly,
132 little is known about ploidy levels in Afromilla, but so far, the two species reported were both
133 polyploids ($2n$ = 64 to 80; octoploid and decaploid; Hjelmqvist 1956; Morton 1993). A recent
134 phylogenomic analysis focused on *Lachemilla* using target enrichment and 32 species of the
135 group detected a high frequency of paralogs shared with Eualchemilla and Afromilla (Morales-
136 Briones et al. 2018b). This paralog frequency suggested a possible ancient WGD event; however,
137 the sampling was limited to one species each of Eualchemilla and Afromilla, and the location
138 and mode of this putative WGD remained uncertain.

139 In this study we sampled 68 species across the major clades of *Alchemilla* s.l., and
140 included 11 additional closely related species in Fragariinae, which allowed us to 1) test for
141 polyploid events in the origin of *Alchemilla* s.l., and 2) explore the reticulate evolution among
142 major clades of *Alchemilla* s.l. using a target enrichment dataset. Given the prevalence of
143 polyploidy and reticulate history within *Alchemilla* s.l., this is an excellent group to explore the

144 utility of tree-based methods for (1) processing paralogs, and (2) detecting and placing WGDs
145 using target enrichment datasets.
146

147 ## MATERIALS AND METHODS

148 ### *Taxon sampling and data collection*

149 We sampled 68 species representing the four major clades of *Alchemilla* s.l. (sensu Gehrke et al.
150 2008), and 11 species to represent all other genera in Fragariinae (except *Chamaecallis*; sensu
151 Dobeš et al. 2015; Morales-Briones and Tank 2019). Additionally, we sampled one species each
152 of *Potentilla*, *Sanguisorba*, and *Rosa* as outgroups. Voucher information is provided in Table S2.
153 We used a Hyb-Seq approach (Weitemier et al. 2014), that combines target enrichment and
154 genome skimming, to capture nuclear exon sequences and off-target cpDNA. We used baits
155 designed for *Fragaria vesca* (strawberry, also a member of Fragariinae) to target 1,419 exons in
156 257 genes (Kamneva et al. 2017). These genes were identified as single-copy orthologs among
157 the apple (*Malus domestica*), peach (*Prunus persica*), and strawberry genomes based on
158 reciprocal nucleotide similarity comparisons. The 257 genes resulted from first retaining only
159 genes >960 bp long and with >85% similarity in pairwise comparisons among the three
160 genomes. The remaining genes were further filtered by removing exons <80 bp long, with GC
161 content <30% or >70%, and with >90% sequence similarity to annotated repetitive DNA in the
162 genome, followed by removing exons with any paralogs with >90% sequence similarity in the
163 same genome (Kamneva et al. 2017).

164 Of the 82 total species, only sequences for *Fragaria vesca*, were from a reference
165 genome (Shulaev et. al 2010). Twenty-two were from a previously published Hyb-Seq dataset
166 using the same bait set as this study (Morales-Briones et al. 2018b; Table S2), including 19

167 species of *Lachemilla* that did not show evidence of hybridization within *Lachemilla,* and one
168 species each of Eualchemilla, Afromilla, and *Aphanes*. Newly generated sequence data for 55
169 species (Table S2) were collected as follows. Total genomic DNA was isolated from silica-dried
170 or herbarium material with a modified CTAB method (Doyle and Doyle 1987). Probe synthesis,
171 library preparation, capture enrichment, and high-throughput sequencing (HiSeq2000 instrument,
172 2 × 101 bp) were carried out at Rapid Genomics LLC (Gainesville, FL, USA). Data for the
173 remaining four species, *Drymocallis glandulosa*, *Potentilla indica*, *Rosa woodsii*, and
174 *Sanguisorba menziesii* were collected as described in Weitemier et al. (2014).
175

176 *Read processing and assembly*

177 We removed sequencing adaptors and trimmed low-quality bases (Phred scores < 20) from raw
178 reads with SeqyClean v.1.10.07 (Zhbannikov et al. 2017) using default settings. Plastomes were
179 assembled using Alignreads v.2.5.2 (Straub et al. 2011) and 12 closely related plastome
180 references (with one Inverted Repeat removed; Table S3). Plastome assemblies were annotated
181 using *Fragaria vesca* as a reference in Geneious v.11.1.5 (Kearse et al. 2012). Assembly of
182 nuclear loci was carried out with HybPiper v.1.3.1 (Johnson et al. 2016) using exons of *F. vesca*
183 as references. Given the large number of paralogs detected in *Lachemilla,* Eualchemilla, and
184 Afromilla, multi-exon gene assemblies resulted in chimeric sequences of exons from distinct
185 paralogs (Morales-Briones et al. 2018b). To avoid chimeric sequences that can affect orthology
186 inference and phylogenetic analyses, assemblies were performed on each exon separately. Only
187 exons with a reference length of ≥ 150 bp were assembled (939 exons from 257 genes). Paralog
188 detection was carried out for all exons with the 'paralog_investigator' option in HybPiper. This
189 option flags loci with potential paralogs when multiple contigs cover at least 85% of the

190 reference sequence length. Exon assemblies that included flagged paralogs were extracted using
191 the 'paralog_retriever' command of HybPiper and used for orthology inference.
192

193 *Orthology inference for nuclear exons*

194 To infer orthologs for phylogenetic analyses, all exons were processed as follows (Fig. 1a).
195 Individual exons were aligned using MACSE v.2.03 (Ranwez et al. 2018) with default
196 parameters. Codons with frameshifts (labeled with '!' by MACSE) were replaced with gaps and
197 aligned columns with more than 90% missing data were removed using Phyx (Brown et al.
198 2017). Initial homolog trees were built using RAxML v.8.2.11 (Stamatakis 2014) with a
199 GTRCAT model and clade support assessed with 100 rapid bootstrap (BS) replicates. Clades and
200 paraphyletic grades that belonged to the same taxon were pruned by keeping only the tip with the
201 highest number of characters in the trimmed alignment following Yang and Smith (2014). To
202 obtain the final homolog trees, outlier tips with unusually long branches were detected and
203 removed by maximally reducing the tree diameter with TreeShrink v.1.3.2 (Mai and Mirarab
204 2018). Orthology inference was carried out using two outgroup-aware strategies from Yang and
205 Smith (2014). We set *Potentilla*, *Sanguisorba*, and *Rosa* as outgroups and all members of
206 Fragariinae as ingroups. First, we used the 'monophyletic outgroup' (MO) approach keeping
207 only ortholog groups with at least 25 ingroup taxa. The MO approach filters for homolog trees
208 with outgroup taxa being monophyletic and single-copy, and therefore filters for single- and low-
209 copy genes. The second approach used was the 'rooted ingroup' (RT), with at least 25 ingroup
210 taxa. The RT approach iteratively searches subtrees of ingroup taxa and cuts them out as rooted
211 trees. Both approaches root the gene tree by the outgroups, traverse the rooted tree from root to
212 tip, and remove the side with fewer taxa (MO) or keep both sides (RT) when gene duplication is

213 detected at any given node. In the case of MO, homolog trees with non-monophyletic outgroups
214 or duplicated taxa in the outgroups are discarded. If no taxon duplication is detected in a
215 homolog tree, the MO approach outputs a one-to-one ortholog. The RT approach maximizes the
216 number of orthologs compared to MO while not requiring monophyletic outgroups and allowing
217 for duplicated taxa in the outgroups but removes outgroups from all orthologs. To add outgroups
218 back to the RT orthologs for downstream analyses, we kept only RT orthologs from homologs
219 that had a MO ortholog (i.e., using only homolog trees with monophyletic and non-duplicated
220 outgroups for both MO and RT). Then we used the outgroups of the MO ortholog for all the RT
221 orthologs of the same homolog (Fig. 1b). Scripts for orthology inference can be found at
222 https://bitbucket.org/dfmoralesb/target_enrichment_orthology.
223
224

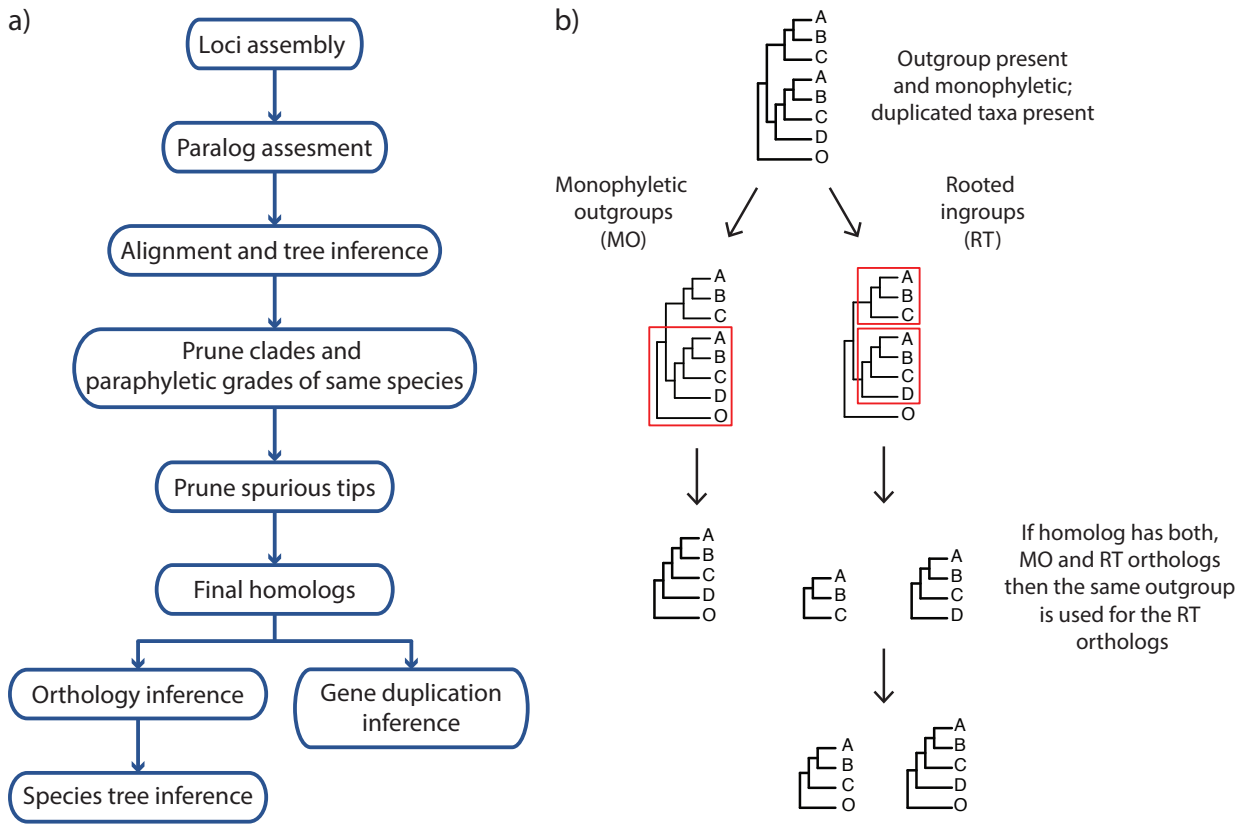

Figure 1. Paralog processing workflow and orthology inference methods used in *Alchemilla* s.l. homolog trees. a) Flow chart of paralog processing and homolog tree inference. b) Only homologs with outgroup present and monophyletic were used for orthology inference. Monophyletic outgroups (MO) will prune 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. Homologs that lack monophyletic outgroups were excluded from further consideration.

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 was aligned using MACSE v.2.03 with default parameters. Codons with frameshifts

239 were replaced with gaps, aligned columns with more than 90% missing data were removed using
240 Phyx, and alignments with at least 150 characters and 25 taxa were retained. We first estimated a
241 maximum likelihood (ML) tree from the concatenated matrices with RAxML using a partition by
242 gene scheme with a GTRGAMMA model for each partition. Clade support was assessed with
243 100 rapid bootstrap (BS) replicates. To estimate a species tree that is statistically consistent with
244 the multi-species coalescent (MSC), we first inferred individual ML gene trees using RAxML
245 with a GTRGAMMA model, and 100 BS replicates to assess clade support. Individual gene trees
246 were then used to estimate a species tree using ASTRAL-III v.5.6.3 (Zhang et al. 2018) using
247 local posterior probabilities (LPP; Sayyari and Mirarab 2016) to assess clade support.

248 To evaluate nuclear gene tree discordance, we calculated the internode certainty all (ICA)
249 value to quantify the degree of conflict on each node of the map tree (e.g., species tree) given
250 individual gene trees (Salichos et al. 2014). Also, we calculated the number of conflicting and
251 concordant bipartitions on each node of the map tree. We calculated both the ICA scores and the
252 number of conflicting/concordant bipartitions with Phyparts (Smith et al. 2015) using individual
253 ortholog trees with BS support of at least 50% for the corresponding node. Additionally, to
254 distinguish strong conflict from weakly supported branches, we evaluated tree conflict and
255 branch support with Quartet Sampling (QS; Pease et al. 2018) using 1,000 replicates. Quartet
256 Sampling subsamples quartets from the input map tree (e.g., species tree) and concatenated
257 alignment to assess the confidence, consistency, and informativeness of each internal branch by
258 the relative frequency of the three possible quartet topologies at each node (Pease et al. 2018).

259 In addition to species tree construction using inferred orthologs, we used a recently
260 developed quartet-based species tree method (ASTRAL-Pro; Zhang et al. 2020a) to estimate the
261 phylogeny of *Alchemilla* s.l. ASTRAL-Pro directly uses multi-labeled gene trees while

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

For the plastome phylogenetic analyses, 74 partial plastome assemblies and eight reference plastome sequences were included (Table S3). Contiguous plastome sequences were aligned using the default settings in MAFFT v.7.307 (Katoh and Standley 2013) and aligned columns with more than 70% missing data were removed with Phyx. We estimated an ML tree of the plastome alignment with RAxML using a partition by coding (CDS) and noncoding regions (introns and intergenic spacers) scheme, with a GTRGAMMA model for each partition and clade support assessed with 100 rapid BS replicates and QS using 1,000 replicates, to detect potential within-plastome conflict in the backbone of *Alchemilla* s.l. as recently reported in other groups (e.g., Gonçalves et al. 2019; Walker et al. 2019; Zhang et al. 2020b; Morales-Briones et al. 2021).

Mapping whole genome duplications

We took two alternative approaches for detecting WGD events by mapping gene duplication events from gene trees onto a map tree (e.g., species tree). The first approach begins by extracting orthogroups from the final homolog trees. Orthogroups are rooted ingroup lineages separated by outgroups that include the complete set of genes in a lineage from a single copy in their common ancestor. We extracted orthogroups requiring at least 50 out of 79 species in

285 Fragariinae. Gene duplication events were then recorded on the most recent common ancestor
286 (MRCA) on the map tree when two or more species overlapped between the two daughter clades
287 Each node on a map tree can be counted only once from each gene tree to avoid nested gene
288 duplications inflating the number of recorded duplications (Yang et al. 2018;
289 https://bitbucket.org/blackrim/clustering, 'extract_clades.py' and 'map_dups_mrca.py'). We
290 mapped duplication events onto both the MO and RT trees using orthogroups from all 923 final
291 homologs, filtering orthogroups using an average BS of at least 50%. We carried out the
292 mapping using two sets of orthogroups, one from all homologs, and the from the longest
293 homologs (the single longest aligned exon per gene) to avoid inflating the counts in multi-exon
294 genes.

295 For the second strategy of WGD mapping, we explicitly tested for polyploidy mode using
296 GRAMPA (Thomas et al. 2017). GRAMPA uses MRCA reconciliation with multi-labeled gene
297 trees to compare allo-or autopolyploid scenarios in singly- or multi-labeled map trees (e.g.,
298 species tree). To reduce the computational burden of searching all possible reconciliations, we
299 constrained searches to only among crown nodes of major clades of *Alchemilla* s.l., which all are
300 well supported (including the 'dissected' and 'lobed' clades of Eualchemilla; see results) and
301 genera within Fragariinae. We ran reconciliation searches using all 923 final homologs, as well
302 as using only the longest homologs (the single longest aligned exon per gene; 256), against either
303 the MO or RT tree. We expected multiple WGD events within *Alchemilla* s.l. (see results), but
304 GRAMPA can only infer one WGD at a time. To disentangle nested duplication events, we also
305 carried out similar GRAMPA reconciliations using the MO tree and sequentially excluding
306 major groups of *Alchemilla* s.l. that were identified as a polyploid clade. We only used the MO
307 tree as it differs from the RT tree only by the location of the 'lobed' clade, which was the first

308 clade identified as allopolyploid (see results) and was removed for subsequent GRAMPA
309 analyses. Finally, to test for a polyploid origin of *Alchemilla* s.l., we carried out searches among
310 the constrained crown node of *Alchemilla* s.l. and the rest of the genera within Fragariinae using
311 the MO and cpDNA trees. The backbone of Fragariinae differed between the MO (same as RT)
312 tree and the cpDNA tree. Thus, we tested how this affected the inference of the polyploid origin
313 of *Alchemilla* s.l. We also carried out similar searches but using each of the five major clades
314 individually.

315 Both approaches used here to detect WGD events use final homolog trees and as any
316 other tree-based method they may be sensitive to tree informativeness. To explore node support
317 across homologs, we run a conflict analysis with Phyparts using individual final homologs trees
318 with a BS support filter of at least 50% for each node. We used both the MO and RT trees as
319 map trees and ran the analysis using all homolog exons as well as only the longest homolog exon
320 per gene.

321

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

323 To obtain further evidence for WGD events and compare them to those inferred from gene
324 duplication events from target enrichment, we analyzed the distribution of synonymous distances
325 (Ks) from RNA-seq data of four species of *Alchemilla* s.l. and nine species of Fragariinae (Table
326 S4). Read processing and transcriptome assembly followed Morales-Briones et al. (2020). For
327 each of the four species, a Ks plot of within-species paralog pairs based on BLASTP hits was
328 done following Yang et al. (2018; https://bitbucket.org/blackrim/clustering; 'ks_plots.py'). Ks
329 peaks were identified using a mixture model as implemented in mixtools v.1.2.0 (Benaglia et al.
330 2009). The optimal number of mixing components was estimated using parametric bootstrap

replicates of the likelihood ratio test statistic (McLachlan and Peel 2000). We tested up to five components using 500 bootstrap replicates in mixtools. Additionally, we used between-species Ks plots to determine the relative timing of the split between two species and compare it to that of WGD events inferred with within-species Ks plots. Ks plots of between-species also followed Yang et al. (2018; 'ks_between_taxa_cds.py'). Lastly, we also attempted to build Ks plots using raw homologs from target enrichment, but the relatively low number of genes (256) failed to produce a meaningful distribution (not shown).

RESULTS

Assembly and orthology inference

The number of assembled exons per species (with > 75% of the target length) ranged from 632 (*Alchemilla fissa*) to 934 (*Dasiphora fruticosa*) out of 939 single-copy exon references from *F. vesca*, with an average of 873 exons (Table S5). The number of exons with paralog warnings ranged from 10 in *Drymocallis glandulosa* to 746 in *Alchemilla mollis* (Table S5). The number of exon alignments with ≥ 25 species was 923 from 256 genes. The orthology inference resulted in 914 MO orthologs (Table S6), and 1,906 RT orthologs (Table S6). The trimmed alignments of the MO orthologs ranged from 136 to 5,740 characters with a mean of 425 characters (median = 268). The concatenated alignment of the MO orthologs, with at least 150 aligned characters and 25 species for each exon, included 910 exons and 387,042 characters with a matrix occupancy of 66%. The trimmed alignments of the RT orthologs ranged from 136 to 5,740 characters with a mean of 394 characters (median = 259). The concatenated alignment of the RT orthologs, with at least 150 aligned characters and 25 species, included 1,894 exons and 746,562 characters with a

353 matrix occupancy of 54%. The chloroplast alignment included 124,079 characters with a matrix
354 occupancy of 77%.
355
356 *Nuclear phylogenetic analyses*
357 All nuclear analyses recovered the monophyly of *Alchemilla* s.l. with maximum support (i.e.,
358 bootstrap percentage [BS] = 100, local posterior probabilities [LPP] = 1.0; Fig. 2; Fig. S1), most
359 informative gene trees being concordant with this node (858 out of 863 for MO; 977/984 for RT;
360 912/932 for ASTRAL-Pro; ICA = 0.95), and full QS support (1.0/–/1.0; i.e., all sampled quartets
361 supported that node). Five major clades were identified within *Alchemilla* s.l.: Afromilla,
362 *Aphanes*, Eualchemilla-'dissected', Eualchemilla-'lobed,' and *Lachemilla*. Moreover, the
363 relationships among these clades showed high levels of discordance and varied among the MO
364 and RT trees.
365 Analyses of the MO orthologs using ASTRAL and concatenated ML approaches resulted
366 in similar topologies for the backbone of *Alchemilla* s.l. (Fig. 2). The monophyly of the five
367 major clades each received maximum support (BS = 100; LPP = 1.0) and had most trees being
368 concordant (except for the two clades of Eualchemilla). Eualchemilla was paraphyletic and split
369 into the 'dissected' and 'lobed' clades. Monophyly of the 'dissected' clade was supported by 118
370 out of 429 informative trees (ICA = 0.08) and strong QS score (0.87/0.34/1), while the 'lobed'
371 clade was supported by 73 out of 420 informative trees (ICA = 0.06) and strong QS score
372 (0.61/0.98/0.99). In both cases, the 'dissected' and 'lobed' clades had a relatively small
373 percentage of supporting trees, but the conflict analysis and QS score did not reveal any well-
374 supported alternative topology. *Aphanes* was recovered as sister to the Eualchemilla-'lobed'
375 clade with relatively low support (BS = 90, LPP = 0.62), 60 concordant trees (out of 430

informative gene trees; ICA = 0.08), and weak QS score (0.016/0.95/0.98) with similar frequencies for the two discordant alternative topologies. The Eualchemilla-‘dissected’ clade was recovered as sister to Eualchemilla-‘lobed’ + *Aphanes* with maximum support, 279 concordant trees (out of 482 informative gene trees; ICA = 0.29), and full QS score. Afromilla was recovered as sister to the clade consisted of Eualchemilla (‘dissected and ‘lobed’) and *Aphanes* with high to low support (BS = 100, LPP = 0.88), only 146 concordant trees (out of 413 informative gene trees; ICA = 0.22), and weak QS support (0.2/0.44/0.99) with a skew in discordance suggesting a possible alternative topology (*Lachemilla* sister to Eualchemilla + *Aphanes*). Lastly, *Lachemilla* was recovered as the sister to the rest of *Alchemilla* s.l.

Analysis of the RT orthologs using ASTRAL and concatenated ML approaches both recovered the same major clades, but they differed in the relationship among these five clades (Fig. 2a; Fig. S1). In both analyses, *Lachemilla*, Afromilla, and *Aphanes* had maximum support (BS = 100; LPP = 1.0) and had most trees being concordant. Eualchemilla was recovered as monophyletic and composed of the 'dissected' and 'lobed' clades. The monophyly of Eualchemilla had high to low support (BS = 99, LPP = 0.63), only 231 concordant trees (out of 819 informative gene trees; ICA = 0.12), and weak QS support (0.023/0.87/0.98) with similar frequencies for the two discordant alternative topologies. Similar to the MO analyses, the ‘dissected’ and ‘lobed’ clades each had low number of concordant trees (218 out of 557 [ICA = 0.19] and 136 out of 707 [ICA = 0.08], respectively), and strong QS support (0.98/0/1 and 0.62/0.17/0.99, respectively). Eualchemilla was recovered as sister of *Aphanes* with maximum support (BS = 100), 348 concordant trees (out of 728 informative trees; ICA = 0.29) and full QS support. The ML concatenated tree (Fig. 2a; Fig. S1) placed Afromilla as sister to the clade formed of Eualchemilla and *Aphanes* with maximum support (BS = 100), 212 concordant gene

trees (out of 771 informative trees; ICA = 0.27), and weak QS support (0.18/0.66/0.99) with no significant skew between the two discordant alternatives. *Lachemilla* was placed as sister to the rest of *Alchemilla* s.l. The ASTRAL tree in turn (Fig. S1a), retrieved *Lachemilla* as sister to the clade formed of Eualchemilla and *Aphanes* with no support (LPP = 0.01), 247 concordant trees (out of 953 informative trees; ICA = 0.19), and QS counter-support (-0.21/0.29/0.99), showing that the majority of the quartets supported one alternative topology (Afromilla sister to Eualchemilla + *Aphanes*). In this case, Afromilla was placed as sister to the rest of *Alchemilla* s.l.

The ASTRAL-Pro analysis using multi-labeled homolog trees recovered the same backbone topology of *Alchemilla* s.l. as the concatenated ML analysis from the RT orthologs (Fig. 2b; Figs S2–S3). All five major clades had the maximum support (LPP = 1.0). Eualchemilla, composed of the 'dissected' and 'lobed' clades, had moderate support (LPP = 0.76) and only 415 concordant trees (out of 1106 informative trees; ICA = 0.17). The ‘dissected’ and ‘lobed’ clades had low numbers of concordant trees (379 out of 941 [ICA = 0.23] and 65 out of 824 [ICA = 0.09], respectively), but did not show signal of any alternative topology. *Aphanes* was placed as the sister of Eualchemilla with maximum support (LPP = 1.0), 426 concordant trees (out of 952 trees; ICA = 0.34), and no support for any major alternative topology. Afromilla was recovered as sister to the clade formed of Eualchemilla and *Aphanes* with low support (LPP = 0.52), 492 concordant trees (out of 953 trees; ICA = 0.42), and no support for any 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’ and ‘lobed’ clades had strong support (BS = 75, QS = 0.8/0.43/0.88 and BS = 100, QS = 0.95/0.25/0.92, respectively). *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.

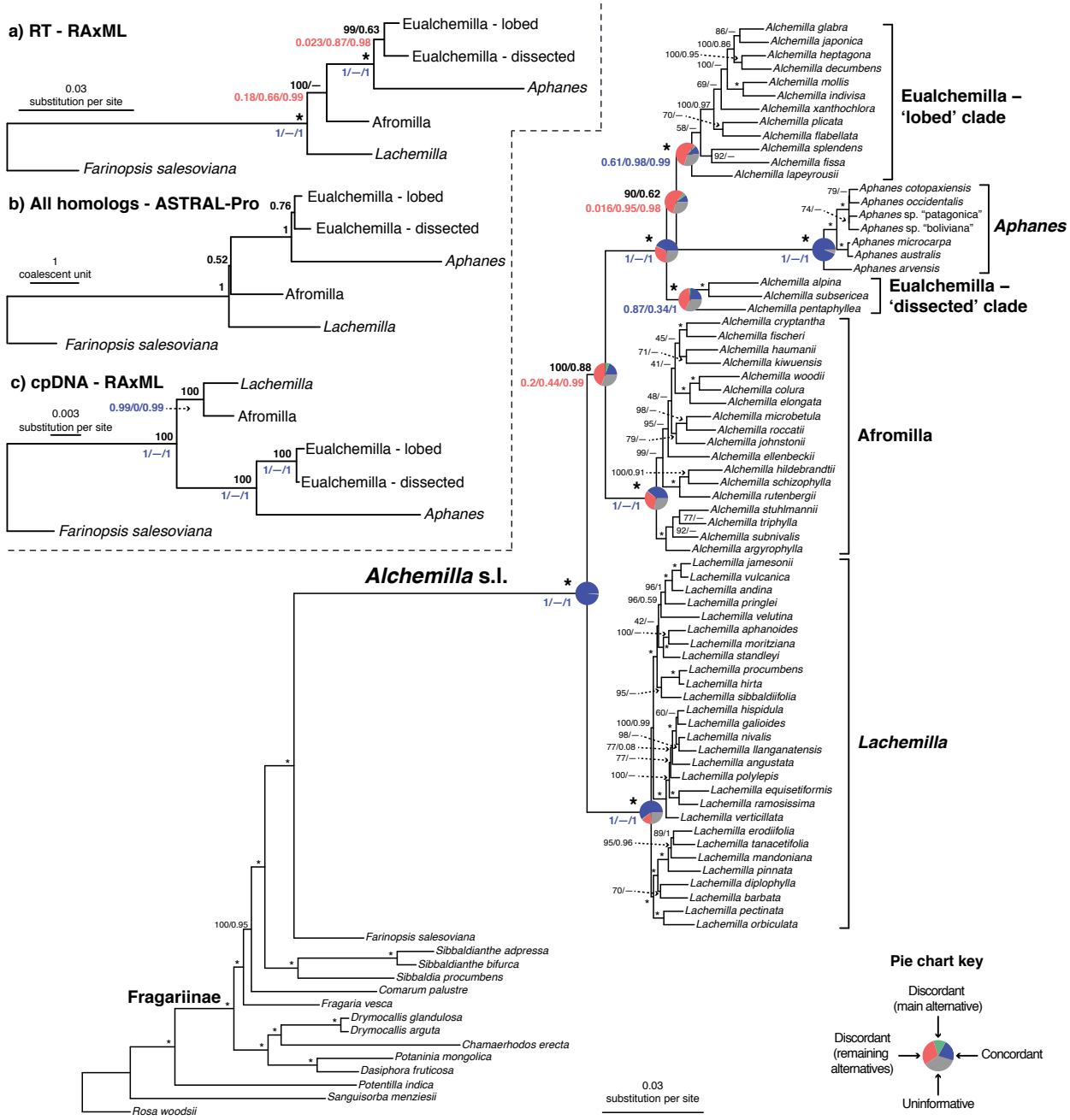

Figure 2. Maximum likelihood phylogeny of *Alchemilla* s.l. inferred from RAxML analysis of the concatenated 910-nuclear exon supermatrix from the 'monophyletic outgroup' (MO) orthologs. Bootstrap support (BS) and Local posterior probability (LLP) are shown above branches. Nodes with full support (BS= 100/LLP= 1) are noted with an asterisk (*). Em dashes (—) denoted alternative topology compared to the ASTRAL 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

436 informativeness. Pie charts for major clades represent the proportion of exon ortholog trees that
437 support that clade (blue), the proportion that support the main alternative bifurcation (green), the
438 proportion that support the remaining alternatives (red), and the proportion (conflict or support)
439 that have < 50% bootstrap support (gray). Gene trees with missing data that were uninformative
440 for the node were ignored. Branch lengths are in number of substitutions per site (scale bar on
441 the bottom). Inset: a) Summary Maximum likelihood phylogeny inferred from RAxML analysis
442 of the concatenated 1,894-nuclear exon supermatrix from the 'rooted ingroup' orthologs (RT).
443 BS and LLP are shown above branches and QS scores below the branches. Branch lengths are in
444 number of substitutions per site; b) Summary ASTRAL-Pro tree inferred from 923 multi-labeled
445 exon homolog trees. LLP are shown next to nodes. Branch lengths are in coalescent units. c)
446 Summary Maximum likelihood phylogeny inferred from RAxML analysis of concatenated
447 partial plastomes. BS and LLP are shown above branches and QS scores below the branches.
448 Branch lengths are in number of substitutions per site.
449

450 *Mapping whole genome duplications*

451 By mapping the most recent common ancestor (MRCA) of gene duplication events from
452 orthogroup trees onto the MO and RT trees, we found four nodes in *Alchemilla* s.l. that each had
453 an elevated proportion of gene duplications (Fig. 3a–b). This trend was consistent regardless of
454 using all 923 homolog exons (868 after orthogroup inference and BS filtering) or using only the
455 256 longest homolog exons per gene (250 after orthogroup inference and BS filtering; Fig. S5).
456 Therefore, here we describe the results only for the latter. These four clades include (Fig. 3a; Fig.
457 S5): 1) the MRCA of *Alchemilla* s.l. (86.0% of the 250 genes show evidence of duplication), 2)
458 the MRCA of Eualchemilla , *Aphanes*, and Afromilla (34.4%), 3) the MRCA of Eualchemilla +
459 *Aphanes* (MO: 18.4%; RT:15.6%), and 4) the MRCA of *Lachemilla* (18.4%). These four nodes
460 have an elevated proportion of gene duplications compared to all other nodes in Fragariinae (Fig.
461 3b) and it is consistent with the number of paralogs counted from the final homolog trees (after
462 pruning of clades or paraphyletic grades of same species; Fig. 3c). Interestingly, although deeply

nested in *Alchemilla* s.l., *Aphanes* had a lower number of paralogs than the rest of *Alchemilla* s.l., resembling the other members of Fragariinae (Fig. 3c).

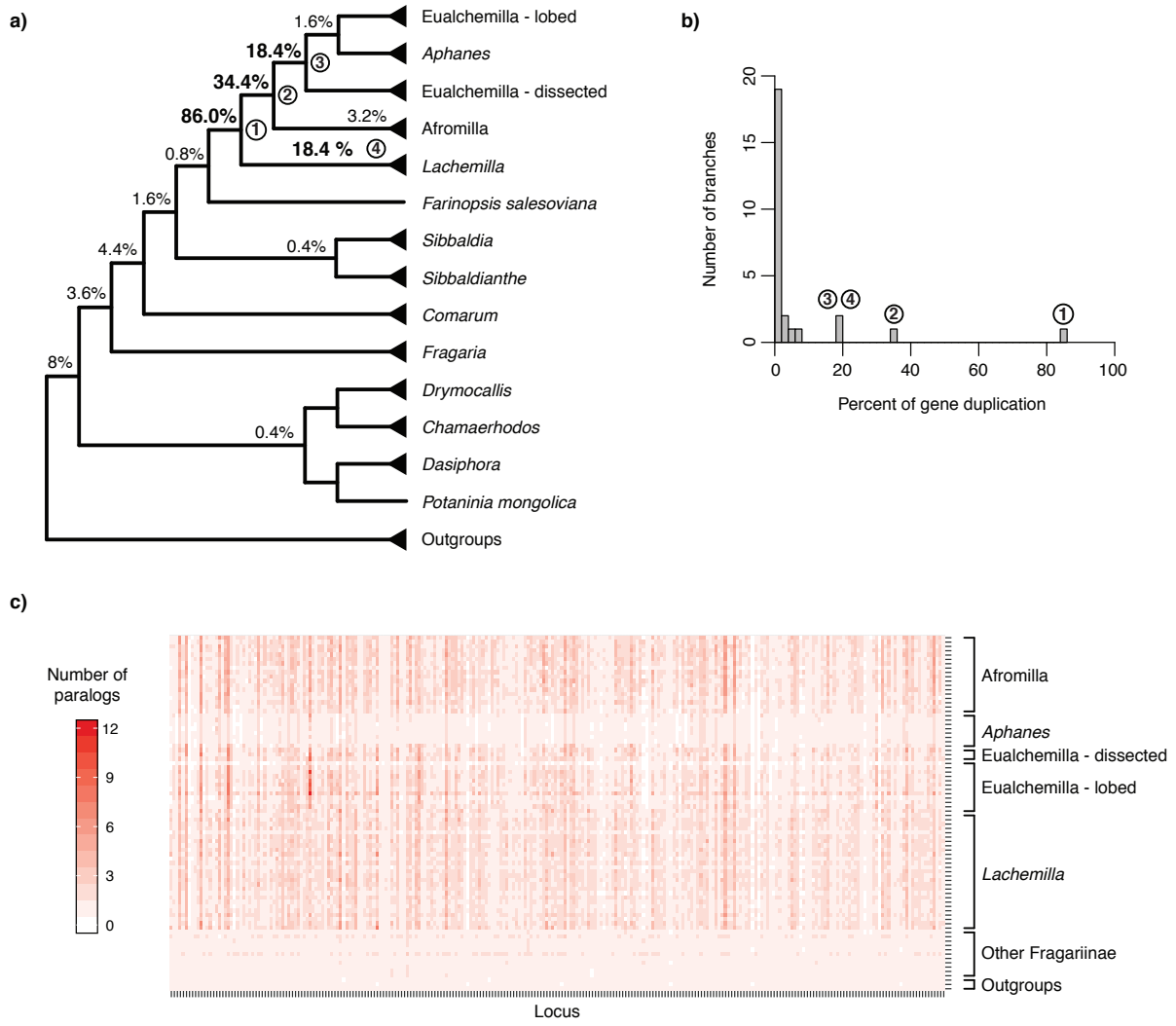

Figure 3. Orthogroup gene duplication mapping results. a) Summarized cladogram of *Alchemilla* s.l. from the 'monophyletic outgroup' (MO) ortholog tree. Percentages next to nodes denote the proportion of duplicated genes when using orthogroups from the longest homologs (250 after orthogroup inference and filtering). Nodes with elevated proportions of gene duplications are numbered 1–4 as referenced in the main text. See Fig. S5 for the full tree. b) Histogram of percentages of gene duplication per branch. c) Number of paralogs per taxa in the final homolog trees. In final homologs clades and paraphyletic grades of the same species were pruned leaving

474 only one tip per species. Each locus is represented by the longest homolog (the single longest
475 aligned exon per gene).
476

477 Bootstrap support for exon homologs were informative (BS ≥ 50%) at most nodes,
478 especially regarding the relationship among the major clades of *Alchemilla* s.l. (Fig. S6).
479 Therefore, uninformative homolog trees were unlikely to affect the results from WGD detection
480 analysis overall. The proportion of uninformative nodes (BS < 50%) were at most 30% in the
481 worst case (Eualchemilla + *Aphanes* + Afromilla) when using all homolog exons. This
482 proportion reduces significantly when using only the longest homolog exons (Fig. S6).
483 Similar to the results of MRCA mapping, the GRAMPA analyses recovered the same
484 results when using all 923 homologs or only the longest homologs (256). GRAMPA
485 reconciliations using all major clades of *Alchemilla* s.l. recovered optimal multi-labeled trees
486 with the best score (i.e., lowest reconciliation score; Fig. S7) where the 'lobed' clade of
487 Eualchemilla was of an allopolyploid origin, but the putative parental lineages varied between
488 the MO and RT trees. The reconciliations using the MO tree (reconciliation score [RS] = 70,250;
489 Fig. 4a; Fig. S8) showed that the 'lobed' clade was of allopolyploid origin between an unsampled
490 or extinct lineage sister to *Aphanes* and an unsampled or extinct lineage (‘lineage’ for short
491 hereafter) sister to 'dissected' + *Aphanes*. In turn, the reconciliations using the RT tree (RS =
492 70,721; Fig. S8) showed that the 'lobed' clade was of allopolyploid origin between a ‘lineage’
493 sister to the 'dissected' clade, and also a ‘lineage’ sister to 'dissected' + *Aphanes*. Alternative
494 multi-labeled trees had higher (worse) RSs (70,482 for MO and 70,739 for RT; Fig. S7). The
495 GRAMPA reconciliations performed on the MO tree with removal of major clades of *Alchemilla*
496 s.l. inferred as allopolyploid resulted in the identification of additional polyploidy events (Fig.
497 4b–d). First, we removed the 'lobed' clade, and this resulted in the recovery of Afromilla as an

allopolyploid clade (RS = 127,836). Afromilla parental lineages were a ‘lineage’ sister to *Aphanes* + the 'dissected' clade, and a ‘lineage’ sister to all remaining *Alchemilla* s.l. (Fig. 4b). Alternative multi-labeled trees reconciliations had scores starting at 127,869 (Fig. S7). The further removal of Afromilla resulted in recovery of the 'dissected' clade as allopolyploid (RS = 167,545). The ‘dissected’ clade had as parental lineages the ‘lineage’ sister to *Aphanes* and the ‘lineage’ sister to all remaining *Alchemilla* s.l. except for *Lachemilla* (Fig. 4c). Other reconciliation alternatives had scores starting at 167,612 (Fig. S7). Finally, the removal of the ‘dissected’ clade resulted in the *Lachemilla* being recovered also as an allopolyploid clade (RS = 181,302). The parental lineages of *Lachemilla* were a ‘lineage’ sister to *Aphanes* and a ‘lineage’ sister to all remaining *Alchemilla* s.l. (Fig. 4d). Alternative multi-labeled trees reconciliations had scores starting at 181,564 (Fig. S7).

The GRAMPA results from the analyses with constrained searches on the crown node of *Alchemilla* s.l. recovered different modes of polyploidy when using the MO tree or the cpDNA tree. The MO tree had *Farinopsis*, *Sibbaldianthe* + *Sibbaldia*, *Comarum*, and *Fragaria* forming a grade sister to *Alchemilla* s.l., while *Drymocallis*, *Chamaerhodos*, *Potaninia*, and *Dasiphora* form a clade that is sister to all other Fragariinae (Fig. 2). The reconciliations using the MO tree resulted in an allopolyploid event for the clade composed of *Alchemilla* s.l., *Farinopsis, Sibbaldianthe*, and *Sibbaldia* (RS = 339,755; Fig. S9). The parental lineages of this clade were a ‘lineage’ sister to *Comarum*, and a ‘lineage’ sister to the grade formed of *Comarum* and *Fragaria* (Fig. S9). Alternative multi-labeled trees had scores starting at 340,053 (Fig. S7). The reconciliations using individual major clades of *Alchemilla* s.l. resulted in identical patterns as in the full constrained analysis (Fig. S10). The cpDNA tree had *Alchemilla* s.l. as part of a grade formed along with *Farinopsis*, *Comarum*, and *Sibbaldianthe* + *Sibbaldia*, while *Fragaria* was

521 recovered as sister to the clade composed of *Drymocallis*, *Chamaerhodos*, *Potaninia*, and
522 *Dasiphora,* which is sister to all other Fragariinae (Fig. S4). The reconciliations on the cpDNA
523 tree recovered *Alchemilla* s.l. as of autopolyploid origin (RS = 364,594; Fig. 4e; Fig. S9).
524 Alternative multi-labeled trees had scores starting at 363,987 (Fig. S7). The analyses using
525 individual major clades of *Alchemilla* s.l. recovered identical patterns as in the full constrained
526 analysis, except for *Aphanes* that resulted in a singly-labeled tree (Fig. S10).

527 To further explore WGD events using alternative data sources, we analyzed Ks plots
528 from genomes and transcriptomes across Fragariinae. The distribution of synonymous distances
529 in the transcriptomes of four species of Eualchemilla (one 'dissected' and three 'lobed') shared
530 three optimal mixing components with a Ks mean at approximately 0.1, 0.34, and 1.67,
531 respectively (Fig. S11). The first two components partially overlapped and corresponded to at
532 least two WGD events in all four sampled species of Eualchemilla, that happened before the
533 splits between the lobed vs. the dissected clades of Eualchemilla (Ks ~ 0.02; Fig. S12). The third
534 shared component corresponds to a whole genome triplication event early in the core eudicots
535 (Jiao et al. 2012; Fig. S11). All nine species from other genera in Fragariinae had two optimal
536 mixing components. One component is a Ks peak at 1.61–1.78 corresponding to the whole
537 genome triplication event early in eudicot (Fig. S11). In the case of the diploid species, the
538 second component represents a small and very young (~ 0.05) peak, most likely the product of
539 small-scale recent gene duplications. The only two polyploid species from the other genera in
540 Fragariinae, *Comarum palustre* (2n=28–64) and *Sibbaldianthe bifurca* (2n=28), had a single
541 additional significant component at 0.11 and 0.08, respectively (Fig. S11). The Ks plots between
542 species of Eualchemilla and Fragariinae species outside of *Alchemilla* s.l., and between species
543 of Fragariinae showed that the WGD events detected in Eualchemilla were not shared with other

genera outside of *Alchemilla* s.l. Likewise, the WGD events in *Comarum palustre* and *Sibbaldianthe bifurca* occurred after the split of the two species (Fig. S12).

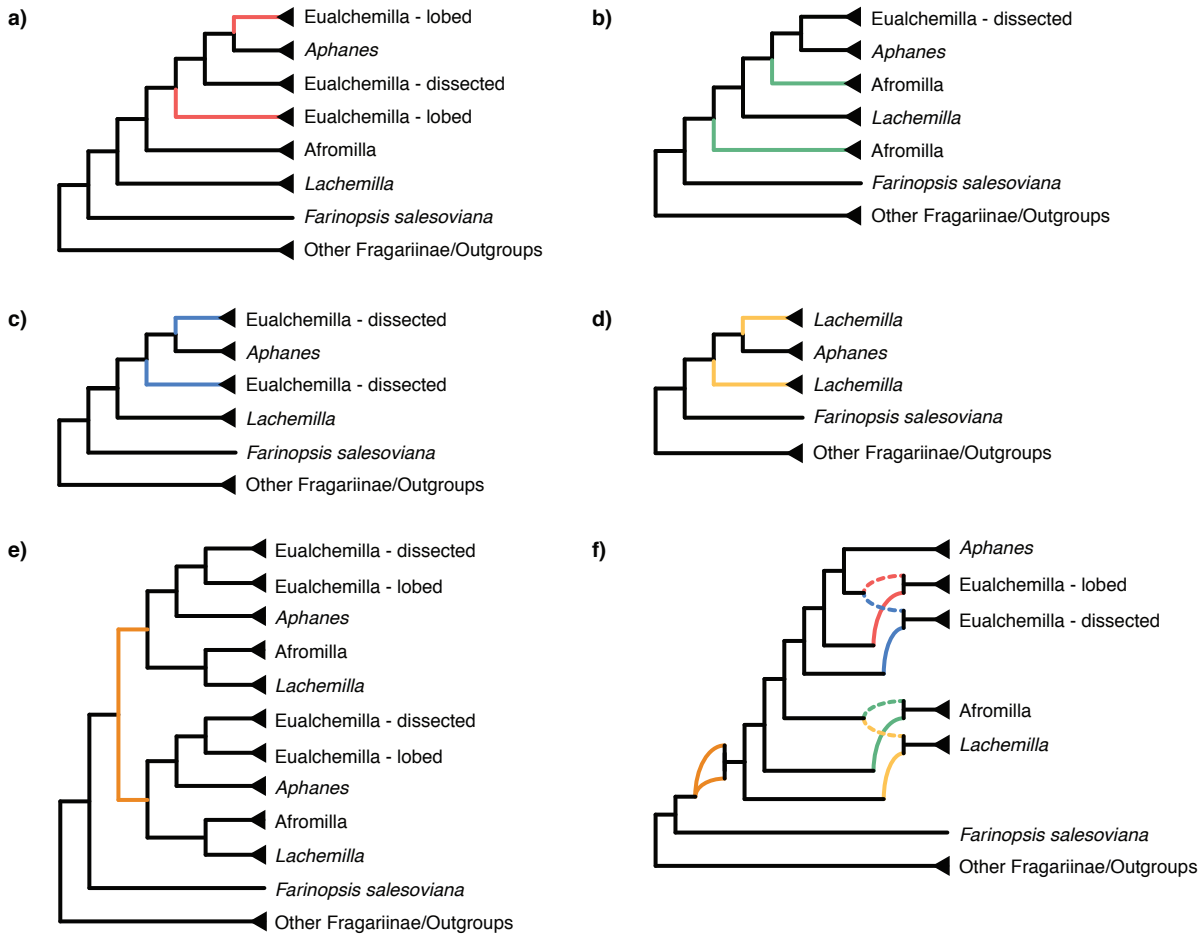

Figure 4. Summary of optimal multi-labeled tree (MUL-tree) inferred from GRAMPA analyses. a) MUL-tree based on reconciling homologs against the species tree inferred from 'monophyletic outgroup' (MO) orthologs including all taxa. Red branches denote the allopolyploid origin of the 'lobed' clade of Eualchemilla. b) MUL-tree after removing the 'lobed' clade of Eualchemilla as in a). Green branches denote the allopolyploid origin of Afromilla. c) MUL-tree after removing Afromilla as in b). Blue branches denote the allopolyploid origin of the 'dissected' clade of Eualchemilla. d). MUL-tree after further removing the 'dissected' clade as in c). Yellow lines denote the allopolyploid origin of *Lachemilla*. e) MUL-tree using constrained searches of the crown node of *Alchemilla* s.l. on the cpDNA tree. Orange branches denote the autopolyploid

557 origin *Alchemilla* s.l. f) Putative summary network of all reticulation events in *Alchemilla* s.l.
558 Colored curved branches denote different polyploid events as in a–e. Dashed curved lines
559 represent the maternal lineage (cpDNA) in allopolyploid events.
560

561 **DISCUSSION**

562 *Processing paralogs in target enrichment datasets*

563 The increased use of target enrichment methods in combination with reduced sequencing costs
564 and higher read coverage have facilitated the recovery of paralogs in such datasets. Paralogy is
565 sometimes viewed as a nuisance for phylogenetic reconstruction and is commonly aimed to be
566 reduced in early stages of experimental design, by targeting only single- or low-copy genes
567 during the selection of loci (e.g., Chamala et al. 2015; Nicholls et al. 2015; Gardner et al. 2016;
568 Kamneva et al. 2017). Still, the recovery of paralogs is inevitable when working with groups
569 where WGD is prevalent, especially in plants, leading to various strategies to remove them prior
570 to phylogenetic analyses. Commonly used target enrichment assembly pipelines (e.g., Faircloth
571 2016; Johnson et al. 2016; Andermann et al. 2018) use different criteria to flag assembled loci
572 with putative paralogs that are later filtered or processed prior to phylogenetic analysis. The most
573 used common approach for dealing with paralogous loci in target enrichment datasets is
574 removing the entire locus that show any signal of potential paralogy (e.g., Crowl et al. 2017;
575 Montes et al. 2019; Bagley et al. 2020). The removal of paralogous loci can significantly reduce
576 the size of target enrichment datasets and most often do not take in consideration the reason why
577 a locus was flagged for putative paralogy (i.e., allelic variation or gene duplication). Orthology
578 inference should be carried for all loci in target enrichment data, as relying on settings in
579 assembly pipelines does not guarantee that non-removed or non-flagged loci are orthologous.
580 Furthermore, removing paralogs before phylogenetic inference eliminates valuable information

581 that could have been used to detect and place WGD events using target enrichment data. Other
582 approaches either retain or remove contigs based on the distinction being putative allelic
583 variation (flagged sequences from monophyletic conspecific groups) or putative paralogs
584 (paralogs from the same species are non-monophyletic) in combination with study-specific
585 threshold or random selection (e.g., Villaverde et al; 2018; Liu et al. 2019; Stubbs et al. 2019), or
586 manual processing (e.g., Garcia et al. 2017; Karimi et al. 2019). As dataset size increases,
587 manual processing becomes prohibitive.

588 The presence of WGDs also poses some challenges for locus assembly. Target
589 enrichment design commonly includes multi-contig targets that assembly pipelines attempt to
590 assemble into single contigs (e.g., Faircloth 2016) or 'supercontigs' composed of multiple exons
591 and partially assembled introns (e.g., Johnson et al. 2016). In groups like *Alchemilla* s.l., where
592 multiple, nested WGD events led to a prevalence of paralogs, 'supercontigs' can produce
593 chimeric assemblies (Morales-Briones et al. 2018b). Instead, we assembled the exons
594 individually to minimize chimeric loci, at the cost of working with some short exons that
595 contribute little phylogenetic information, which can affect orthology inference and downstream
596 analyses. Therefore, it is important to take this into consideration during target enrichment
597 experimental design, and to preferentially target long exons when possible in groups where
598 WGD is expected. An alternative strategy to avoid chimeric supercontigs when gene duplications
599 are frequent is to perform a preliminary orthology inference in single exon-based trees and then
600 use the inferred orthologs as a reference to reassemble the loci into 'supercontigs' (e.g., Gardner
601 et al. 2020; Karimi et al. 2020). Another aspect to take in consideration during or right after
602 assembly is allele phasing. While phasing heterozygous loci, from population or individual
603 variation, have been shown to have minimal impact in phylogenetic reconstruction in target

604 enrichment data (e.g., Kates et al. 2018), the effect on unphased or merged loci in cases of WGD
605 can be larger and be a source of gene tree error. Here we were interested in ancient WGD in
606 *Alchemilla* s.l. and relied on enough sequencing coverage and sequence dissimilarity to assemble
607 separate paralogs (homoeologs in the case of allopolyploidy) that can be flagged as such by
608 HybPiper. While we obtained a large number of deep paralogs across *Alchemilla* s.l. (Fig. 3c;
609 Table S5), there is still the possibility of some locus included merged sequences from paralogs
610 with high sequence similarity. Paralog merger should be more problematic in cases of recent
611 allopolyploidy or neo-allopolyploidy taxa. To this end, recently developed tools have been
612 designed to phase gene copies into polyploid subgenomes using phylogenetic and similarity
613 approaches (e.g., Freyman et al. 2020, Nauheimer et al. 2020).

614 The utility of paralogs for phylogenetic reconstruction in target enrichment datasets is
615 gaining more attention (e.g., Johnson et al. 2016; Gardner et al. 2020). A few studies have
616 considered tree-based orthology inference to process affected loci (e.g., Garcia et al. 2017;
617 Moore et al. 2018, Morales-Briones et al. 2018b), but in some cases the orthology approaches
618 used cannot be applied to other groups. Here we demonstrated the utility of automated, tree-
619 based orthology inference methods (Yang and Smith 2014), originally designed for genomic or
620 transcriptomic datasets, to infer orthology from paralog-flagged loci in a target enrichment
621 dataset. Our approach facilitates the automated inference of orthologs while maximizing the
622 number of loci retained for downstream analyses. These methods are agnostic of the data source
623 and should work for any type of target enrichment dataset (e.g., anchored phylogenomics, exon
624 capture; Hyb-Seq, ultraconserved elements).

625 Orthology inference methods used here (Yang and Smith et al. 2014) are a powerful tool
626 for target enrichment datasets. In the case of allopolyploidy, however, these methods can

627 introduce bias in the distribution of ortholog trees inferred. In the case of MO, each time a gene
628 duplication event is detected, the side with a smaller number of taxa is removed. When
629 allopolyploidy occurs, MO may bias towards one subgenome due to 1) bias in gene loss between
630 subgenomes. Even if the submissive subgenome is present in some parts of the genome, it is
631 differentially lost in a higher number of loci, 2) bias in bait design. In the case of *Alchemilla* s.l.,
632 this is less likely as baits are designed in outgroups. If baits are designed according to ingroup
633 taxa, depending on which taxa were used it can have higher affinity to one subgenome instead of
634 another, and 3) unequal sampling of parental lineages. If one parental lineage is more densely
635 sampled than the other or one parental lineage is unsampled, the two subgenomes will be in
636 species-rich versus species-poor clades respectively in gene trees. One could alternatively
637 preserve a random side each time a gene duplication event is identified. However, in practice, the
638 side with a smaller number of taxa often contains misassembled or misplaced sequences. The RT
639 method of separating duplicated gene copies, on the other hand, keeps any subtree with sufficient
640 number of taxa, but removes outgroups, and worked best when hierarchical outgroups were
641 included in the taxon sampling. Therefore, both MO and RT lose information, especially in cases
642 with complex, nested polyploidy. Recently developed methods based on quartet similarity
643 (Zhang et al. 2020a) or Robinson-Foulds distances (Molloy and Warnow 2020) can directly
644 estimate species trees from multi-labeled trees that are consistent with the MSC and gene
645 duplication and loss without inferring orthologs (for a recent review see Smith and Hahn 2020).
646 However, their behavior on complex datasets using archival materials is yet to be explored. For
647 example, both methods do not define ingroup-outgroup relationships a priori, and correctly
648 inferring the root of homolog trees can be challenging with missing data, or when WGD occurs
649 near the root. In addition, none of these above species tree reconstruction methods (Molloy and

650 Warnow 2020; Zhang et al. 2020a) were designed to handle reticulate relationships. This can
651 result in species tree topology that is an “average” between subgenomes. Depending on the
652 topological distance of subgenomes, the resulting species tree may not represent any subgenome
653 history. Finally, most current methods for evaluating node support still require orthologous gene
654 trees as input. In such cases tools like Phyparts can still be used to visualize gene tree
655 discordance and calculate ICA scores using multi-labeled trees.
656

657 *Phylogenetic implications in Alchemilla s.l.*

658 Previous phylogenetic studies established the monophyly *Alchemilla* s.l. and four major clades of
659 the group (Gehrke et al. 2008; 2016), but the relationship among them and the placement of
660 *Alchemilla* s.l. within Fragariinae remain unresolved. Our nuclear and plastid analyses both
661 confirmed the monophyly of *Alchemilla* s.l. and its sister relationship to *Farinopsis*, as
662 previously shown by Morales-Briones and Tank (2019) based on plastome sequences only.
663 Gehrke et al. (2008) identified two well supported clades within Eualchemilla that were
664 distinguished by leaf shape, namely the 'dissected' and 'lobed' clades. Most species of
665 Eualchemilla have a leaf shape consistent with their clade placement, but some had different leaf
666 shapes that were attributed to their hybrid origin between the two clades (Gehrke et al. 2008).
667 More recently, Gehrke et al. (2016) and Morales-Briones and Tank (2019) found that
668 Eualchemilla is not monophyletic in analyses that included the external transcribed spacer (ETS)
669 of the nuclear ribosomal DNA (nrDNA) cistron. Both studies found *Aphanes* nested between the
670 'dissected' and 'lobed' clades of Eualchemilla. Our analyses of the nuclear loci supported the
671 monophyly of 'dissected' and 'lobed' clades, but the monophyly of Eualchemilla had low support
672 (Fig. 2; Fig. S1). The analysis using only the MO orthologs even weakly supported the 'lobed'

673 clade as sister of *Aphanes* (Fig. 2). In contrast, our plastome analysis recovered a well-supported,
674 monophyletic Eualchemilla, as well as well-supported 'dissected' and 'lobed' clades. Both nuclear
675 and plastid analyses strongly supported the clade composed of *Aphanes* and both clades of
676 Eualchemilla (Fig. 2; Fig. S4), a relationship that is consistent with previous nuclear and plastid
677 analyses (Gehrke et al. 2008; 2016; Morales-Briones and Tank 2019). Given the revealed
678 hybridization in the evolution and early divergence within *Alchemilla* s.l., the non-monophyly of
679 Eualchemilla could be explained by ancient gene flow or an allopolyploid origin of the
680 'dissected' and 'lobed' clades (Fig. 4a,c; see below). Besides the well supported relationship of
681 Eualchemilla + *Aphanes*, our nuclear analysis showed high levels of conflict among other major
682 clades in *Alchemilla* s.l. (Fig. 2; Fig. S1) which could also be explained by additional ancient
683 allopolyploid events (Fig. 4; see below).
684

685 *Ancient polyploidy in Alchemilla s.l.*

686 Whole-genome duplications are frequent across Rosaceae (Dickinson et al. 2007; Xiang et al.
687 2017), and allopolyploidy has been suggested as the primary source for the cytonuclear
688 discordance in Fragariinae (Lundberg et al. 2009; Gehrke et al. 2016; Morales-Briones and Tank
689 2019). We recovered four nodes in *Alchemilla* s.l. with a high percentage of gene duplications
690 (Fig. 3a; Fig. S5). One of the nodes showing a high percentage of gene duplication (18.4%) was
691 the MRCA of *Aphanes* and both clades of Eualchemilla (node 3 in Fig. 3a; Fig. S5). This
692 duplication event agreed with the MRCA of the ancestral lineages inferred with GRAMPA for
693 the allopolyploid origin of the 'lobed' clade of Eualchemilla (Fig. 4a). Moreover, the GRAMPA
694 reconciliations after the removal of the 'lobed' clade and Afromilla inferred a scenario where the
695 'dissected' clade is of allopolyploid origin with one of the parental lineages as sister to *Aphanes*

696 (Fig. 4c). Although there is some uncertainty about the placement of the parental lineage of
697 'dissected' clade, due to the removal of major clades for the GRAMPA analyses, the cpDNA tree
698 suggest that it is likely sister to the parental lineage of ‘lobed’ clade that is also sister to *Aphanes*.
699 Ks plots of all species of the 'dissected' and 'lobed' clades had two peaks that are not shared with
700 members of Fragariinae (Figs S11–S12), suggesting that at least two WGD events have
701 happened between the stem lineage of *Alchemilla* s.l. to the crown node of the ‘dissected’ and the
702 ‘lobed’ clades. The between-species Ks plots between 'dissected' and 'lobed' (Fig. S12), showed
703 that the split between these two groups is more recent than the WGD events, suggesting a single
704 origin (or very close in time) of both clades. Still, the sister relationship of the 'dissected' and
705 'lobed' clades is not supported by nuclear genes, suggesting that the two clades of Eualchemilla
706 might had independent allopolyploid origins, while sharing the same or a closely related
707 maternal lineage (cpDNA; Fig. 4f).

708 The GRAMPA reconciliation, after the removal of the ‘lobed’ clade, recovered an
709 allopolyploid origin of Afromilla (Fig. 4b) with a MRCA of the ancestral lineages at the crown
710 of the remaining *Alchemilla* s.l. Similarly, the further removal of both Afromilla and the
711 ‘dissected’ clade recovered *Lachemilla* as allopolyploid, with the MRCA of parental lineages
712 mapped to the crown of the remaining *Alchemilla* s.l. (Fig. 4d). In the case of *Lachemilla,*
713 because of the removal of major clades for the GRAMPA analyses, there is also no certainty in
714 the placement of its parental lineages. Still, Afromilla and *Lachemilla* are sisters in the cpDNA
715 tree (Fig 1C.), suggesting these two share the same or a closely related maternal lineage (Fig. 4f).
716 The high percentage of gene duplication (34.4%) placed at the MRCA of the clade composed of
717 Afromilla, Eualchemilla, and *Aphanes* (node 2 in Fig. 3a), could be explained in part by the
718 allopolyploid origin of Afromilla.

719 Finally, the node with the highest percentage of duplicated genes (86%) was placed at the
720 MRCA of *Alchemilla* s.l. (node 1 in Fig. 3a). The GRAMPA analysis using the MO tree showed
721 an allopolyploid event for the clade that included *Alchemilla* s.l., *Farinopsis salesoviana*,
722 *Sibbaldia*, and *Sibbaldianthe* (Fig. S9). However, an allopolyploid origin of *Farinopsis*
723 *salesoviana*, *Sibbaldia*, and *Sibbaldianthe* is not supported by chromosome numbers, orthogroup
724 gene duplication counts, or Ks plots. All members in Fragariinae, with the exception of
725 *Alchemilla* s.l. mainly consists of diploid species and base chromosome number of seven (x = 7),
726 including *Sibbaldia* and *Sibbaldianthe.* On the other hand, *Alchemilla* s.l. has a base number of
727 eight (x = 8) and contains mostly species with high ploidy levels (octoploid to 24-ploid), with the
728 exception of most species of *Aphanes* (2n=16) and one species of *Lachemilla* (*L. mandoniana,*
729 2n=16). Also, our gene duplication counts show low percentages (1.6%) of gene duplication for
730 the MRCA of the GRAMPA-inferred allopolyploid clade or the MRCA (3.6%) of the inferred
731 parental lineages (Fig. 3). Previous phylotranscriptomic analyses of Rosaceae (Xiang et al. 2017)
732 that included one species each of the ‘dissected’ and 'lobed' clades of Eualchemilla, found
733 33.21% of duplicated genes for the MRCA of these two clades, but did not recover any other
734 node with elevated gene duplications within Fragariinae. The Ks plots of the four species of
735 *Alchemilla* s.l. all showed peaks with similar Ks means, but these peaks were not shared with
736 species of *Sibbaldia* and *Sibbadianthe* (Fig. S11). Furthermore, the between-species Ks plots
737 showed that the WGD events detected in *Alchemilla* were more recent than the split with
738 members of Fragariinae (Fig. S12). Although the chromosome number and Ks data for
739 *Farinopsis salesoviana* are not available, all the above evidence suggest an unlikely
740 allopolyploid origin of the clade consisting of *Farinopsis, Sibbaldia*, *Sibadianthe*, and *Alchemilla*
741 sl. On the other hand, the GRAMPA reconciliations using the cpDNA tree resulted in an optimal

742 multi-labeled tree where *Alchemilla* s.l. had an autopolyploid origin (Fig. 4e). This scenario is
743 compatible with the high percentage of gene duplication at the MRCA of *Alchemilla* s.l. and the
744 low percentage of gene duplication in the backbone of the rest of Fragariinae. Another
745 compatible scenario is an allopolyploid origin of *Alchemilla* s.l. where both parental lineages are
746 missing or extinct, but this scenario is indistinguishable from autopolyploidy. The atypical high
747 proportion of gene duplication at the base of *Alchemilla* s.l. can be explained by the
748 autopolyploid event at this branch. In addition, given the short branch lengths among major
749 clades within *Alchemilla* s.l., gene tree estimation error (e.g., uninformative genes), incomplete
750 lineage sorting (ILS), allopolyploid events among major clades of *Alchemilla* s.l., and/or
751 homoeologous exchanges among subgenomes (Edger et al. 2018; McKain et al. 2018) can all
752 contribute to additional gene duplication events mapped to the MRCA of *Alchemilla* s.l.

753 Although deeply nested in *Alchemilla* s.l., remarkably, *Aphanes* showed a significantly
754 lower number of paralogs than the rest of *Alchemilla* s.l. (Fig. 3). The relatively low number of
755 paralogs, its diploid species being mainly diploid, and the best GRAMPA reconciliation resulting
756 in a singly-labeled tree (Fig. S10), suggesting that *Aphanes* is a functional diploid clade. One
757 plausible scenario is that post-polyploid diploidization (reviewed in Mandáková and Lysak 2018)
758 occurred after the autopolyploidy event at the base of *Alchemilla* s.l. After diploidization,
759 Afromilla, Eualchemilla ('lobed' and 'dissected' clades), and *Lachemilla* originated from
760 allopolyploid events (Fig. 4f). On the other hand, *Aphanes* seems to descend from a diploidized
761 ancestor that did not duplicate further. The orthogroup gene duplication mapping showed
762 *Aphanes* as part of a clade that had nested high proportions of gene duplication in the orthogroup
763 mapping (Fig. 3a-b, nodes 2–3). But this does not necessarily mean that *Aphanes* should show
764 the same duplication pattern, or neither does it contradict its diploid condition, as a duplication

765 event does not affect or include all descendants of the mapped MRCA in the map tree (e.g.,
766 species trees).

767 GRAMPA has been shown to be useful to identify multiple polyploidy events in the same
768 tree (e.g., Thomas et al. 2017; Guo et al. 2020; Koenen et al. 2020), but a tree-based approach
769 can also be sensitive to gene tree estimation error or ILS (Thomas et al. 2017). Methods to infer
770 species networks in the presence of ILS (e.g., Solís-Lemus and Ané 2016; Wen et al. 2018) could
771 also be used to explore the prevalence of ancient hybridization in *Alchemilla* s.l. Although these
772 methods are under constant development and improvement, they are still only tractable in simple
773 scenarios with few reticulation events (Hejase and Liu 2016; Kamneva and Rosenberg 2017).
774 Similarly, the signal of the *D*-Statistic (Green et al. 2010; Durand et al. 2011), commonly used to
775 detect introgression, can be lost or distorted in presence of multiple reticulations (Elworth et al.
776 2018). Complex reticulate scenarios like *Alchemilla* s.l. are likely to face these problems and
777 have phylogenetic network and *D*-statistic identifiability issues as seen in other groups (e.g.,
778 Morales-Briones et al. 2021).

779

780 *Conclusions*

781 In this study, we have shown the utility of target enrichment datasets in combination with
782 tree-based methods for orthology inference and WGD investigation. Here, we used *Alchemilla*
783 s.l. to highlight the importance of processing paralogs, rather than discarding them before
784 phylogenetic analysis, to shed light on the complex polyploidy histories. We showed evidence
785 that the entire *Alchemilla* s.l. is the product of an ancient autopolyploidy event, and that
786 Afromilla, Eualchemilla ('lobed' and 'dissected' clades), and *Lachemilla* originated from
787 subsequent and nested ancient allopolyploid events. Our results from analyzing target enrichment

788 data corroborated with previously published chromosome numbers and distribution of Ks values
789 from transcriptomes. Our analyses has several important implications for future target
790 enrichment projects, including 1) design baits to obtain a relatively large number of loci as this is
791 required for accurate species tree and networks estimation in complex scenarios (e.g., higher
792 levels of ILS; Solís-Lemus and Ané 2016; Nute et al. 2018), 2) increase the length of individual
793 loci to improve the information content of individual gene trees for proper tree-based orthology
794 inference and identifying gene duplication events, and 3) design baits to minimize lineage-
795 specific and paralog-specific capture efficiency and missing data. Furthermore, in target
796 enrichment, unlike genome or transcriptome data, only a few hundreds of genes are typically
797 recovered with levels of missing data that varies by lineage and are non-random. This limits the
798 utility of target enrichment for generating Ks plots, and creates the need to carefully scrutinize
799 the variation in percentage of gene duplications among nodes. In the end, even with these
800 limitations, target enrichment is an overall valuable and cost-effective approach of genomic
801 subsampling to explore patterns of reticulation and WGD, especially in groups for which whole
802 genome or transcriptome data are not possible to generate, including from museum/herbarium
803 specimens. As research continues to deepen in other clades across the Tree of Life using similar
804 target enrichment methods, we expect that other complex patterns of duplication and reticulation,
805 as those shown here in *Alchemilla* s.l. will continue to emerge.
806

807 **SUPPLEMENTARY MATERIAL**

808 Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/.[NNNN]
809
810

ACKNOWLEDGMENTS

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

REFERENCES

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

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

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

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

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

833 Buddenhagen C., Lemmon A.R., Lemmon E.M., Bruhl J., Cappa J., Clement W.L., Donoghue
834 M.J., Edwards E.J., Hipp A.L., Kortyna M., Mitchell N., Moore A., Prychid C.J.,
835 Segovia-Salcedo M.C., Simmons M.P., Soltis P.S., Wanke S., Mast A. 2016. Anchored
836 Phylogenomics of Angiosperms I: Assessing the Robustness of Phylogenetic Estimates.
837 bioRxiv.:086298.
838 Chamala S., García N., Godden G.T., Krishnakumar V., Jordon-Thaden I.E., De Smet R.,
839 Barbazuk W.B., Soltis D.E., Soltis P.S. 2015. MarkerMiner 1.0: A new application for
840 phylogenetic marker development using angiosperm transcriptomes. Appl. Plant Sci.
841 3:1400115.
842 Crowl A.A., Manos P.S., McVay J.D., Lemmon A.R., Lemmon E.M., Hipp A.L. 2019.
843 Uncovering the genomic signature of ancient introgression between white oak lineages
844 (Quercus). New Phytol.:nph.15842.
845 Dickinson T.A., Lo E., Talent N. 2007. Polyploidy, reproductive biology, and Rosaceae:
846 understanding evolution and making classifications. Plant Syst. Evol. 266:59–78.
847 Dobeš C., Lückl A., Kausche L., Scheffknecht S., Prohaska D., Sykora C., Paule J. 2015. Parallel
848 origins of apomixis in two diverged evolutionary lineages in tribe Potentilleae
849 (Rosaceae): Origin of Apomixis in Potentilleae. Bot. J. Linn. Soc. 177:214–229.
850 Doyle J.J., Doyle J.L. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf
851 tissue. Phytochem. Bull. 19:11–15.
852 Dunn C.W., Howison M., Zapata F. 2013. Agalma: an automated phylogenomics workflow.
853 BMC Bioinformatics. 14:330.
854 Durand E.Y., Patterson N., Reich D., Slatkin M. 2011. Testing for Ancient Admixture between
855 Closely Related Populations. Mol. Biol. Evol. 28:2239–2252.

856 Edger P.P., McKain M.R., Bird K.A., VanBuren R. 2018. Subgenome assignment in
857 allopolyploids: challenges and future directions. Curr. Opin. Plant Biol. 42:76–80.
858 Elworth R.A.L., Allen C., Benedict T., Dulworth P., Nakhleh L.K. 2018. DGEN: A Test Statistic
859 for Detection of General Introgression Scenarios. WABI.
860 Emms D.M., Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative
861 genomics. Genome Biol. 20:238.
862 Eriksson T., Lundberg M., Töpel M., Östensson P., Smedmark J.E.E. 2015. Sibbaldia: a
863 molecular phylogenetic study of a remarkably polyphyletic genus in Rosaceae. Plant
864 Syst. Evol. 301:171–184.
865 Faircloth B.C. 2016. PHYLUCE is a software package for the analysis of conserved genomic
866 loci. Bioinforma. Oxf. Engl. 32:786–788.
867 Fernández R., Gabaldon T., Dessimoz C. 2020. Orthology: Definitions, Prediction, and Impact
868 on Species Phylogeny Inference. In: Scornavacca C., Delsuc F., Galtier N., editors.
869 Phylogenetics in the Genomic Era. No commercial publisher | Authors open access
870 book. p. 2.4:1--2.4:14.
871 Fitch W.M. 1970. Distinguishing Homologous from Analogous Proteins. Syst. Biol. 19:99–113.
872 Forrest L.L., Hart M.L., Hughes M., Wilson H.P., Chung K.-F., Tseng Y.-H., Kidner C.A. 2019.
873 The Limits of Hyb-Seq for Herbarium Specimens: Impact of Preservation Techniques.
874 Front. Ecol. Evol. 7:439.
875 Freyman W.A., Johnson M.G., Rothfels C.J. 2020. homologizer: Phylogenetic phasing of gene
876 copies into polyploid subgenomes. bioRxiv 2020.10.22.351486
877 García N., Folk R.A., Meerow A.W., Chamala S., Gitzendanner M.A., Oliveira R.S. de, Soltis
878 D.E., Soltis P.S. 2017. Deep reticulation and incomplete lineage sorting obscure the

879 diploid phylogeny of rain-lilies and allies (Amaryllidaceae tribe Hippeastreae). Mol.
880 Phylogenet. Evol. 111:231–247.
881 Gardner E.M., Johnson M.G., Pereira J.T., Ahmad Puad A.S., Arifiani D., Sahromi, Wickett
882 N.J., Zerega N.J.C. 2020. Paralogs and off-target sequences improve phylogenetic
883 resolution in a densely-sampled study of the breadfruit genus (Artocarpus, Moraceae).
884 Syst. Bio. syaa073.
885 Gardner E.M., Johnson M.G., Ragone D., Wickett N.J., Zerega N.J.C. 2016. Low-coverage,
886 whole-genome sequencing of Artocarpus camansi (Moraceae) for phylogenetic marker
887 development and gene discovery. Appl. Plant Sci. 4:1600017.
888 Gehrke B., Bräuchler C., Romoleroux K., Lundberg M., Heubl G., Eriksson T. 2008. Molecular
889 phylogenetics of Alchemilla, Aphanes and Lachemilla (Rosaceae) inferred from plastid
890 and nuclear intron and spacer DNA sequences, with comments on generic classification.
891 Mol. Phylogenet. Evol. 47:1030–1044.
892 Gehrke B., Kandziora M., Pirie M.D. 2016. The evolution of dwarf shrubs in alpine
893 environments: a case study of Alchemilla in Africa. Ann. Bot. 117:121–131.
894 Glover N., Dessimoz C., Ebersberger I., Forslund S.K., Gabaldón T., Huerta-Cepas J., Martin
895 M.-J., Muffato M., Patricio M., Pereira C., da Silva A.S., Wang Y., Sonnhammer E.,
896 Thomas P.D. 2019. Advances and Applications in the Quest for Orthologs. Mol. Biol.
897 Evol. 36:2157–2164.
898 Gonçalves D.J.P., Simpson B.B., Ortiz E.M., Shimizu G.H., Jansen R.K. 2019. Incongruence
899 between gene trees and species trees and phylogenetic signal variation in plastid genes.
900 Mol. Phylogenet. Evol. 138:219–232.

901 Green R.E., Krause J., Briggs A.W., Maricic T., Stenzel U., Kircher M., Patterson N., Li H., Zhai
902 W., Fritz M.H.Y., Hansen N.F., Durand E.Y., Malaspinas A.S., Jensen J.D., Marques-
903 Bonet T., Alkan C., Prufer K., Meyer M., Burbano H.A., Good J.M., Schultz R., Aximu-
904 Petri A., Butthof A., Hober B., Hoffner B., Siegemund M., Weihmann A., Nusbaum C.,
905 Lander E.S., Russ C., Novod N., Affourtit J., Egholm M., Verna C., Rudan P., Brajkovic
906 D., Kucan Z., Gusic I., Doronichev V.B., Golovanova L.V., Lalueza-Fox C., de la
907 Rasilla M., Fortea J., Rosas A., Schmitz R.W., Johnson P.L.F., Eichler E.E., Falush D.,
908 Birney E., Mullikin J.C., Slatkin M., Nielsen R., Kelso J., Lachmann M., Reich D.,
909 Paabo S. 2010. A Draft Sequence of the Neandertal Genome. Science. 328:710–722.
910 Guo X., Mandáková T., Trachtová K., Özüdoğru B., Liu J., Lysak M.A. 2020. Linked by
911 ancestral bonds: multiple whole-genome duplications and reticulate evolution in a
912 Brassicaceae tribe. Mol. Biol. Evol. msaa327.
913 Hayirhoğlu-Ayaz S., İnceer H., Frost-Olsen P. 2006. Chromosome counts in the genus
914 *Alchemilla* (Rosaceae) from SW Europe. Folia Geobot. 41:335–344.
915 Hejase H.A., Liu K.J. 2016. A scalability study of phylogenetic network inference methods using
916 empirical datasets and simulations involving a single reticulation. BMC Bioinformatics.
917 17:422.
918 Hjelmqvist H. 1956. The embryology of some African *Alchemilla* species. Bot. Not. 109:21–32.
919 Huang C.-H., Zhang C., Liu M., Hu Y., Gao T., Qi J., Ma H. 2016. Multiple Polyploidization
920 Events across Asteraceae with Two Nested Events in the Early History Revealed by
921 Nuclear Phylogenomics. Mol. Biol. Evol. 33:2820–2835.
922 Izmailow R. Karyological studies in species of *Alchemilla* L. from the series *Calycinae* Bus.
923 (section *Brevicaulon* Rothm.). Acta Biol. Cracoviensia Ser. Bot. 23:117–130.

Jiao Y., Wickett N.J., Ayyampalayam S., Chanderbali A.S., Landherr L., Ralph P.E., Tomsho L.P., Hu Y., Liang H., Soltis P.S., Soltis D.E., Clifton S.W., Schlarbaum S.E., Schuster S.C., Ma H., Leebens-Mack J., dePamphilis C.W. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature. 473:97–100.

Jiao Y., Leebens-Mack J., Ayyampalayam S., Bowers J.E., McKain M.R., McNeal J., Rolf M., Ruzicka D.R., Wafula E., Wickett N.J., Wu X., Zhang Y., Wang J., Zhang Y., Carpenter E.J., Deyholos M.K., Kutchan T.M., Chanderbali A.S., Soltis P.S., Stevenson D.W., McCombie R., Pires J., Wong G., Soltis D.E., dePamphilis C.W. 2012. A genome triplication associated with early diversification of the core eudicots. Genome Biol. 13:R3.

Johnson M.G., Gardner E.M., Liu Y., Medina R., Goffinet B., Shaw A.J., Zerega N.J.C., Wickett N.J. 2016. HybPiper: Extracting Coding Sequence and Introns for Phylogenetics from High-Throughput Sequencing Reads Using Target Enrichment. Appl. Plant Sci. 4:1600016.

Jones K.E., Fér T., Schmickl R.E., Dikow R.B., Funk V.A., Herrando-Moraira S., Johnston P.R., Kilian N., Siniscalchi C.M., Susanna A., Slovák M., Thapa R., Watson L.E., Mandel J.R. 2019. An empirical assessment of a single family-wide hybrid capture locus set at multiple evolutionary timescales in Asteraceae. Appl. Plant Sci. 7:e11295.

Kamneva O.K., Rosenberg N.A. 2017. Simulation-Based Evaluation of Hybridization Network Reconstruction Methods in the Presence of Incomplete Lineage Sorting. Evol. Bioinforma. 13:117693431769193.

945 Kamneva O.K., Syring J., Liston A., Rosenberg N.A. 2017. Evaluating allopolyploid origins in
946 strawberries (Fragaria) using haplotypes generated from target capture sequencing.
947 BMC Evol. Biol. 17:401.
948 Karimi N., Grover C.E., Gallagher J.P., Wendel J.F., Ané C., Baum D.A. 2020. Reticulate
949 Evolution Helps Explain Apparent Homoplasy in Floral Biology and Pollination in
950 Baobabs (Adansonia; Bombacoideae; Malvaceae). Syst. Biol. 69:462–478.
951 Katoh K., Standley D.M. 2013. MAFFT Multiple Sequence Alignment Software Version 7:
952 Improvements in Performance and Usability. Mol. Biol. Evol. 30:772–780.
953 Kates H.R., Johnson M.G., Gardner E.M., Zerega N.J.C., Wickett N.J. 2018. Allele phasing has
954 minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in
955 a case study of Artocarpus. Am. J. Bot. 105:404–416.
956 Kearse M., Moir R., Wilson A., Stones-Havas S., Cheung M., Sturrock S., Buxton S., Cooper A.,
957 Markowitz S., Duran C., Thierer T., Ashton B., Meintjes P., Drummond A. 2012.
958 Geneious Basic: An integrated and extendable desktop software platform for the
959 organization and analysis of sequence data. Bioinformatics. 28:1647–1649.
960 Kocot K.M., Citarella M.R., Moroz L.L., Halanych K.M. 2013. PhyloTreePruner: A
961 Phylogenetic Tree-Based Approach for Selection of Orthologous Sequences for
962 Phylogenomics. Evol. Bioinforma. Online. 9:429–435.
963 Koenen E.J.M., Ojeda D.I., Bakker F.T., Wieringa J.J., Kidner C., Hardy O.J., Pennington R.T.,
964 Herendeen P.S., Bruneau A., Hughes C.E. 2020. The Origin of the Legumes is a
965 Complex Paleopolyploid Phylogenomic Tangle closely associated with the Cretaceous-
966 Paleogene (K-Pg) Mass Extinction Event. Syst. Biol. syaa041.

967 Larridon I., Villaverde T., Zuntini A.R., Pokorny L., Brewer G.E., Epitawalage N., Fairlie I.,
968 Hahn M., Kim J., Maguilla E., Maurin O., Xanthos M., Hipp A.L., Forest F., Baker W.J.
969 2020. Tackling Rapid Radiations With Targeted Sequencing. Front. Plant Sci. 10:1655.
970 Leebens-Mack J.H., Barker M.S., Carpenter E.J., Deyholos M.K., Gitzendanner M.A., Graham
971 S.W., Grosse I., Li Z., Melkonian M., Mirarab S., Porsch M., Quint M., Rensing S.A.,
972 Soltis D.E., Soltis P.S., Stevenson D.W., Ullrich K.K., Wickett N.J., DeGironimo L.,
973 Edger P.P., Jordon-Thaden I.E., Joya S., Liu T., Melkonian B., Miles N.W., Pokorny L.,
974 Quigley C., Thomas P., Villarreal J.C., Augustin M.M., Barrett M.D., Baucom R.S.,
975 Beerling D.J., Benstein R.M., Biffin E., Brockington S.F., Burge D.O., Burris J.N.,
976 Burris K.P., Burtet-Sarramegna V., Caicedo A.L., Cannon S.B., Çebi Z., Chang Y.,
977 Chater C., Cheeseman J.M., Chen T., Clarke N.D., Clayton H., Covshoff S., Crandall-
978 Stotler B.J., Cross H., dePamphilis C.W., Der J.P., Determann R., Dickson R.C., Di
979 Stilio V.S., Ellis S., Fast E., Feja N., Field K.J., Filatov D.A., Finnegan P.M., Floyd
980 S.K., Fogliani B., García N., Gâteblé G., Godden G.T., Goh F. (Qi Y., Greiner S.,
981 Harkess A., Heaney J.M., Helliwell K.E., Heyduk K., Hibberd J.M., Hodel R.G.J.,
982 Hollingsworth P.M., Johnson M.T.J., Jost R., Joyce B., Kapralov M.V., Kazamia E.,
983 Kellogg E.A., Koch M.A., Von Konrat M., Könyves K., Kutchan T.M., Lam V., Larsson
984 A., Leitch A.R., Lentz R., Li F.-W., Lowe A.J., Ludwig M., Manos P.S., Mavrodiev E.,
985 McCormick M.K., McKain M., McLellan T., McNeal J.R., Miller R.E., Nelson M.N.,
986 Peng Y., Ralph P., Real D., Riggins C.W., Ruhsam M., Sage R.F., Sakai A.K.,
987 Scascitella M., Schilling E.E., Schlösser E.-M., Sederoff H., Servick S., Sessa E.B.,
988 Shaw A.J., Shaw S.W., Sigel E.M., Skema C., Smith A.G., Smithson A., Stewart C.N.,
989 Stinchcombe J.R., Szövényi P., Tate J.A., Tiebel H., Trapnell D., Villegente M., Wang

990 C.-N., Weller S.G., Wenzel M., Weststrand S., Westwood J.H., Whigham D.F., Wu S.,
991 Wulff A.S., Yang Y., Zhu D., Zhuang C., Zuidof J., Chase M.W., Pires J.C., Rothfels
992 C.J., Yu J., Chen C., Chen L., Cheng S., Li J., Li R., Li X., Lu H., Ou Y., Sun X., Tan
993 X., Tang J., Tian Z., Wang F., Wang J., Wei X., Xu X., Yan Z., Yang F., Zhong X.,
994 Zhou F., Zhu Y., Zhang Y., Ayyampalayam S., Barkman T.J., Nguyen N., Matasci N.,
995 Nelson D.R., Sayyari E., Wafula E.K., Walls R.L., Warnow T., An H., Arrigo N.,
996 Baniaga A.E., Galuska S., Jorgensen S.A., Kidder T.I., Kong H., Lu-Irving P., Marx
997 H.E., Qi X., Reardon C.R., Sutherland B.L., Tiley G.P., Welles S.R., Yu R., Zhan S.,
998 Gramzow L., Theißen G., Wong G.K.-S., One Thousand Plant Transcriptomes Initiative.
999 2019. One thousand plant transcriptomes and the phylogenomics of green plants. Nature.
1000 574:679–685.
1001 Li L. 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome
1002 Res. 13:2178–2189.
1003 Li Z., Baniaga A.E., Sessa E.B., Scascitelli M., Graham S.W., Rieseberg L.H., Barker M.S.
1004 2015. Early genome duplications in conifers and other seed plants. Sci. Adv.
1005 1:e1501084.
1006 Liu Y., Johnson M.G., Cox C.J., Medina R., Devos N., Vanderpoorten A., Hedenäs L., Bell N.E.,
1007 Shevock J.R., Aguero B., Quandt D., Wickett N.J., Shaw A.J., Goffinet B. 2019.
1008 Resolution of the ordinal phylogeny of mosses using targeted exons from organellar and
1009 nuclear genomes. Nat. Commun. 10:1485.
1010 Lundberg M., Töpel M., Eriksen B., Nylander J.A.A., Eriksson T. 2009. Allopolyploidy in
1011 Fragariinae (Rosaceae): Comparing four DNA sequence regions, with comments on
1012 classification. Mol. Phylogenet. Evol. 51:269–280.

1013 Lynch M., Conery J.S. 2000. The Evolutionary Fate and Consequences of Duplicate Genes.
1014 Science. 290:1151–1155.
1015 Mai U., Mirarab S. 2018. TreeShrink: fast and accurate detection of outlier long branches in
1016 collections of phylogenetic trees. BMC Genomics. 19:272.
1017 Mandáková T., Lysak M.A. 2018. Post-polyploid diploidization and diversification through
1018 dysploid changes. Curr. Opin. Plant Biol. 42:55–65.
1019 Mandel J.R., Dikow R.B., Funk V.A., Masalia R.R., Staton S.E., Kozik A., Michelmore R.W.,
1020 Rieseberg L.H., Burke J.M. 2014. A Target Enrichment Method for Gathering
1021 Phylogenetic Information from Hundreds of Loci: An Example from the Compositae.
1022 Appl. Plant Sci. 2:1300085.
1023 Nauheimer L., Weigner N., Joyce E., Crayn D., Clarke C., Nargar K. 2020. HybPhaser: a
1024 workflow for the detection and phasing of hybrids in target capture datasets. bioRxiv
1025 2020.10.27.354589
1026 McKain M.R., Estep M.C., Pasquet R., Layton D.J., Vela Díaz D.M., Zhong J., Hodge J.G.,
1027 Malcomber S.T., Chipabika G., Pallangyo B., Kellogg E.A. 2018. Ancestry of the two
1028 subgenomes of maize. bioRxiv.:352351.
1029 McKain M.R., Tang H., McNeal J.R., Ayyampalayam S., Davis J.I., dePamphilis C.W., Givnish
1030 T.J., Pires J.C., Stevenson D.W., Leebens-Mack J.H. 2016. A Phylogenomic Assessment
1031 of Ancient Polyploidy and Genome Evolution across the Poales. Genome Biol. Evol.
1032 8:1150–1164.
1033 McLachlan G., Peel D. 2000. Finite Mixture Models. New York: Wiley.
1034 Molloy E.K., Warnow T. 2020. FastMulRFS: fast and accurate species tree estimation under
1035 generic gene duplication and loss models. Bioinformatics. 36:i57–i65.

1036 Montes J.R., Peláez P., Willyard A., Moreno-Letelier A., Piñero D., Gernandt D.S. 2019.
1037 Phylogenetics of Pinus Subsection Cembroides Engelm. (Pinaceae) Inferred from Low-
1038 Copy Nuclear Gene Sequences. Syst. Bot. 44:501–518.
1039 Montgomery L. 1997. Contributions to a cytological catalogue of the British and Irish flora, 5.
1040 Watsonia. 21:365–368.
1041 Moore A.J., Vos J.M.D., Hancock L.P., Goolsby E., Edwards E.J. 2018. Targeted Enrichment of
1042 Large Gene Families for Phylogenetic Inference: Phylogeny and Molecular Evolution of
1043 Photosynthesis Genes in the Portullugo Clade (Caryophyllales). Syst. Biol. 67:367–383.
1044
1045 Morales-Briones D.F., Romoleroux K., Kolář F., Tank D.C. 2018a. Phylogeny and Evolution of
1046 the Neotropical Radiation of *Lachemilla* (Rosaceae): Uncovering a History of Reticulate
1047 Evolution and Implications for Infrageneric Classification. Syst. Bot. 43:17–34.
1048 Morales-Briones D.F., Liston A., Tank D.C. 2018b. Phylogenomic analyses reveal a deep history
1049 of hybridization and polyploidy in the Neotropical genus *Lachemilla* (Rosaceae). New
1050 Phytol. 218:1668–1684.
1051 Morales-Briones D.F., Tank D.C. 2019. Extensive allopolyploidy in the neotropical genus
1052 *Lachemilla* (Rosaceae) revealed by. Am. J. Bot. 106:415–437.
1053 Morales-Briones D.F., Kadereit G., Tefarikis D.T., Moore M.J., Smith S.A.,
1054 Brockington S.F., Timoneda A., Yim W.C., Cushman J.C., Yang Y. 2021. Disentangling
1055 Sources of Gene Tree Discordance in Phylogenomic Datasets: Testing Ancient
1056 Hybridizations in Amaranthaceae s.l. Syst. Biol. 70:219–235
1057 Morton J. 1993. Chromosome numbers and polyploidy in the flora of Cameroons Mountain.
1058 Opera Bot. 121:159–172.

1059 Nicholls J.A., Pennington R.T., Koenen E.J.M., Hughes C.E., Hearn J., Bunnefeld L., Dexter
1060 K.G., Stone G.N., Kidner C.A. 2015. Using targeted enrichment of nuclear genes to
1061 increase phylogenetic resolution in the neotropical rain forest genus Inga (Leguminosae:
1062 Mimosoideae). Front. Plant Sci. 6:710.
1063 Nute M., Chou J., Molloy E.K., Warnow T. 2018. The performance of coalescent-based species
1064 tree estimation methods under models of missing data. BMC Genomics. 19:286.
1065 Panchy N., Lehti-Shiu M., Shiu S.-H. 2016. Evolution of Gene Duplication in Plants. Plant
1066 Physiol. 171:2294–2316.
1067 Pease J.B., Brown J.W., Walker J.F., Hinchliff C.E., Smith S.A. 2018. Quartet Sampling
1068 distinguishes lack of support from conflicting support in the green plant tree of life. Am.
1069 J. Bot. 105:385–403.
1070 Perry L.M. 1929. A Tentative Revision of Alchemilla § Lachemilla. Contrib. Gray Herb. Harv.
1071 Univ.:1–57.
1072 Ranwez V., Douzery E.J.P., Cambon C., Chantret N., Delsuc F. 2018. MACSE v2: Toolkit for
1073 the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. Mol.
1074 Biol. Evol. 35:2582–2584.
1075 Salichos L., Stamatakis A., Rokas A. 2014. Novel Information Theory-Based Measures for
1076 Quantifying Incongruence among Phylogenetic Trees. Mol. Biol. Evol. 31:1261–1271.
1077 Sayyari E., Mirarab S. 2016. Fast Coalescent-Based Computation of Local Branch Support from
1078 Quartet Frequencies. Mol. Biol. Evol. 33:1654–1668.
1079 Shulaev V., Sargent D.J., Crowhurst R.N., Mockler T.C., Folkerts O., Delcher A.L., Jaiswal P.,
1080 Mockaitis K., Liston A., Mane S.P., Burns P., Davis T.M., Slovin J.P., Bassil N.,
1081 Hellens R.P., Evans C., Harkins T., Kodira C., Desany B., Crasta O.R., Jensen R.V.,

1082 Allan A.C., Michael T.P., Setubal J.C., Celton J.-M., Rees D.J.G., Williams K.P., Holt
1083 S.H., Rojas J.J.R., Chatterjee M., Liu B., Silva H., Meisel L., Adato A., Filichkin S.A.,
1084 Troggio M., Viola R., Ashman T.-L., Wang H., Dharmawardhana P., Elser J., Raja R.,
1085 Priest H.D., Bryant D.W., Fox S.E., Givan S.A., Wilhelm L.J., Naithani S., Christoffels
1086 A., Salama D.Y., Carter J., Girona E.L., Zdepski A., Wang W., Kerstetter R.A., Schwab
1087 W., Korban S.S., Davik J., Monfort A., Denoyes-Rothan B., Arus P., Mittler R., Flinn
1088 B., Aharoni A., Bennetzen J.L., Salzberg S.L., Dickerman A.W., Velasco R.,
1089 Borodovsky M., Veilleux R.E., Folta K.M. 2011. The genome of woodland strawberry
1090 (Fragaria vesca). Nat. Genet. 43:109–116.

1091 Smedmark J.E.E., Eriksson T., Evans R.C., Campbell C.S. 2003. Ancient Allopolyploid
1092 Speciation in Geinae (Rosaceae): Evidence from Nuclear Granule-Bound Starch
1093 Synthase (GBSSI) Gene Sequences. Syst. Biol. 52:374–385.

1094 Smith M.L., Hahn M.W. 2020. New Approaches for Inferring Phylogenies in the Presence of
1095 Paralogs. Trends Genet.

1096 Smith S.A., Moore M.J., Brown J.W., Yang Y. 2015. Analysis of phylogenomic datasets reveals
1097 conflict, concordance, and gene duplications with examples from animals and plants.
1098 BMC Evol. Biol. 15:150.

1099 Soják J. 2008. Notes on Potentilla XXI. A new division of the tribe Potentilleae (Rosaceae) and
1100 notes on generic delimitations. Bot. Jahrb. Für Syst. Pflanzengesch. Pflanzengeogr.
1101 127:349–358.

1102 Solís-Lemus C., Ané C. 2016. Inferring Phylogenetic Networks with Maximum
1103 Pseudolikelihood under Incomplete Lineage Sorting. PLoS Genet. 12:e1005896–21.

1104 Stamatakis A. 2014. RAxML version 8 - a tool for phylogenetic analysis and post-analysis of
1105 large phylogenies. Bioinformatics. 30:1312–1313.
1106 Straub S.C., Fishbein M., Livshultz T., Foster Z., Parks M., Weitemier K., Cronn R.C., Liston A.
1107 2011. Building a model: developing genomic resources for common milkweed
1108 (Asclepias syriaca) with low coverage genome sequencing. BMC Genomics. 12:211.
1109 Stubbs R.L., Folk R.A., Xiang C.-L., Chen S., Soltis D.E., Cellinese N. 2020. A Phylogenomic
1110 Perspective on Evolution and Discordance in the Alpine-Arctic Plant Clade Micranthes
1111 (Saxifragaceae). Front. Plant Sci. 10:1773.
1112 Thomas G.W.C., Ather S.H., Hahn M.W. 2017. Gene-Tree Reconciliation with MUL-Trees to
1113 Resolve Polyploidy Events. Syst. Biol. 66:1007–1018.
1114 Villaverde T., Pokorny L., Olsson S., Rincón-Barrado M., Johnson M.G., Gardner E.M., Wickett
1115 N.J., Molero J., Riina R., Sanmartín I. 2018. Bridging the micro- and macroevolutionary
1116 levels in phylogenomics: Hyb-Seq solves relationships from populations to species and
1117 above. New Phytol. 220:636–650.
1118 Walker J.F., Walker-Hale N., Vargas O.M., Larson D.A., Stull G.W. 2019. Characterizing gene
1119 tree conflict in plastome-inferred phylogenies. PeerJ. 7:e7747.
1120 Walters S., Boznan V. *Alchemilla faeroensis* (Lange) Buser and *A. alpina* L. Proc. Bot. Soc. Br.
1121 Isles. 7:83.
1122 Weitemier K., Straub S.C.K., Cronn R.C., Fishbein M., Schmickl R., McDonnell A., Liston A.
1123 2014. Hyb-Seq: Combining Target Enrichment and Genome Skimming for Plant
1124 Phylogenomics. Appl. Plant Sci. 2:1400042.
1125 Wen D., Yu Y., Zhu J., Nakhleh L. 2018. Inferring Phylogenetic Networks Using PhyloNet.
1126 Syst. Biol. 67:735–740.

1127 Xiang Y., Huang C.-H., Hu Y., Wen J., Li S., Yi T., Chen H., Xiang J., Ma H. 2017. Evolution
1128 of Rosaceae Fruit Types Based on Nuclear Phylogeny in the Context of Geological
1129 Times and Genome Duplication. Mol. Biol. Evol. 34:262–281.
1130 Yang Y., Moore M.J., Brockington S.F., Mikenas J., Olivieri J., Walker J.F., Smith S.A. 2018.
1131 Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy
1132 events in Caryophyllales, including two allopolyploidy events. New Phytol. 217:855–
1133 870.
1134 Yang Y., Moore M.J., Brockington S.F., Soltis D.E., Wong G.K.-S., Carpenter E.J., Zhang Y.,
1135 Chen L., Yan Z., Xie Y., Sage R.F., Covshoff S., Hibberd J.M., Nelson M.N., Smith
1136 S.A. 2015. Dissecting Molecular Evolution in the Highly Diverse Plant Clade
1137 Caryophyllales Using Transcriptome Sequencing. Mol. Biol. Evol. 32:2001–2014.
1138 Yang Y., Smith S.A. 2014. Orthology Inference in Nonmodel Organisms Using Transcriptomes
1139 and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for
1140 Phylogenomics. Mol. Biol. Evol. 31:3081–3092.
1141 Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: polynomial time species tree
1142 reconstruction from partially resolved gene trees. BMC Bioinformatics. 19:153.
1143 Zhang C., Scornavacca C., Molloy E.K., Mirarab S. 2020a. ASTRAL-Pro: Quartet-Based
1144 Species-Tree Inference despite Paralogy. Mol. Biol. Evol. 37:3292–3307.
1145 Zhang R., Wang Y.-H., Jin J.-J., Stull G.W., Bruneau A., Cardoso D., De Queiroz L.P., Moore
1146 M.J., Zhang S.-D., Chen S.-Y., Wang J., Li D.-Z., Yi T.-S. 2020b. Exploration of Plastid
1147 Phylogenomic Conflict Yields New Insights into the Deep Relationships of
1148 Leguminosae. Syst. Biol. 69:613–622.

1149 Zhbannikov I.Y., Hunter S.S., Foster J.A., Settles M.L. 2017. SeqyClean: A Pipeline for High-
1150 throughput Sequence Data Preprocessing. Proc. 8th ACM Int. Conf. Bioinforma.
1151 Comput. Biol. Health Inform.:407–416.

1152