A New Pipeline for Removing Paralogs in Target Enrichment Data

Wenbin Zhou[1*], John Soghigian[2,3], Qiu-Yun (Jenny) Xiang[1*]

1 Department of Plant and Microbial Biology, North Carolina State University, Raleigh, NC 27965, USA

2 Department of Entomology and Plant Pathology, North Carolina State University, Raleigh, NC 27965, USA

3 Present Address: Department of Comparative Biology and Experimental Medicine, Faculty of Veterinary Medicine, University of Calgary, Calgary T2N 4Z6, Alberta, Canada

*Corresponding authors e-mail addresses: jenny_xiang@ncsu.edu; wzhou10@ncsu.edu

Abstract.—Target enrichment (such as Hyb-Seq) is a well-established high throughput sequencing method that has been increasingly used for phylogenomic studies. Unfortunately, current widely used pipelines for analysis of target enrichment data do not have a vigorous procedure to remove paralogs in target enrichment data. In this study, we develop a pipeline we call Putative Paralogs Detection (PPD) to better address putative paralogs from enrichment data. The new pipeline is an add-on to the existing HybPiper pipeline, and the entire pipeline applies criteria in both sequence similarity and heterozygous sites at each locus in the identification of paralogs. Users may adjust the thresholds of sequence identity and heterozygous sites to identify and remove paralogs according to the level of phylogenetic divergence of their group of interest. The new pipeline also removes highly polymorphic sites attributed to errors in sequence assembly and gappy regions in the alignment. We demonstrated the value of the new pipeline using empirical data generated from Hyb-Seq and the Angiosperm 353 kit for two woody genera *Castanea* (Fagaceae, Fagales) and *Hamamelis* (Hamamelidaceae, Saxifragales). Comparisons of datasets showed that the PPD identified many more putative paralogs than the popular method HybPiper. Comparisons of tree topologies and divergence times showed evident differences between data from HybPiper and data from our new PPD pipeline. We further evaluated the accuracy and error rates of PPD by BLAST mapping of putative paralogous and orthologous sequences to a reference genome sequence of *Castanea mollissima*. Compared to HybPiper alone, PPD identified substantially more paralogous gene sequences that mapped to multiple regions of the reference genome (31 genes for PPD compared with 4 genes for HybPiper alone). In conjunction with HybPiper, paralogous genes identified by both pipelines can be removed resulting in the construction of more robust orthologous gene datasets for phylogenomic and divergence time analyses. Our study demonstrates the value of Hyb-Seq with data derived from

the Angiosperm 353 probe set for elucidating species relationships within a genus, and argues for the importance of additional steps to filter paralogous genes and poorly aligned regions (e.g., as occur through assembly errors), such as our new PPD pipeline described in this study.

Keywords: Hyb-Seq, Angiosperm 353, paralogs, phylogenomics, divergence time, *Castanea*, *Hamamelis*

High throughput sequencing (HTS) technologies, such as those associated with amplicon sequencing, restriction site digestion, target enrichment, and transcriptome sequencing, have empowered systematists and evolutionary biologists to infer phylogeny with genome-wide molecular markers for a better understanding of species relationships and to answer evolutionary questions with new perspectives that were not possible in the past (e.g. Pais et al. 2017, 2018; Dong et al. 2019; Fu et al. 2019; One Thousand Plant Transcriptomes Initiative 2019; Du et al. 2020; Gaynor et al. 2020; Zhou et al. 2020; Thomas et al. 2021; see reviews in Lemmon and Lemmon 2013; Dodsworth et al. 2019). Among these HTS technologies, target enrichment (Hyb-Seq in plants or sequence capture - Weitemier et al. 2014; and ultraconserved elements, UCEs, in animals - Faircloth et al. 2012) is highly promising and increasingly used for phylogenomic studies of lineages across different evolutionary timescales (e.g. Faircloth et al. 2013; McCormack et al. 2013; Leache et al. 2015; Léveillé-Bourret et al. 2018; Gaynor et al. 2020). The target enrichment method produces data from a targeted set of highly conserved genomic regions (and their flanking areas), often protein coding genes, using probes designed from prior knowledge of target sequences, either from the organism of interest, or a closely related species. The method is highly valued for its repeatability between experiments and between labs if the same probes are used (Harvey et al. 2016), and for generating a lasting and amplifiable resource for comparative studies at multiple taxonomic scales. Data from target enrichment have been shown to be suitable to phylogenomic studies of both deep and shallow phylogenetic divergence, depending on the probes used, because the data contain both conserved coding sequences and their flanking variable sequences (Lemmon et al. 2012; Faircloth et al. 2013; McCormack et al. 2013; Leache et al. 2015; Barrow et al. 2018; Léveillé-Bourret et al. 2018; Banker et al. 2020; Gaynor et al. 2020).

The development of the Angiosperm 353 kit (Johnson et al. 2019), which captures 353 low copy nuclear genes across angiosperms, has enabled phylogenomic studies across angiosperm lineages from family to genus (e.g., Gaynor et al. 2020 for Diapensaceae; Larridon et al. 2020 for Cyperaceae; Murphy et al. 2020 for *Nepenthes* in Nepenthaceae; Shee et al. 2020 for *Scheffera* in Araliaceae). An explosion of phylogenomic studies using the Angiosperm 353 probes is expected in the plant systematics community in the coming years. This endeavor will result in combinable datasets for building the "tree of life" of angiosperms through global-scale analysis (Dodsworth et al. 2019; Johnson et al. 2019). However, the universal probe kit has a disadvantage compared to taxon-specific kits in that the 353 target genes may or may not all be single copy across all species on which the kit is used, and probe binding affinity may cause probes to target unintended paralogous sequences (McCartney et al. 2016). In other words, the potential high divergence of some of the 353 target genes among the diverse angiosperm genomes poses a concern on possible prevalence of paralogs in the Hyb-Seq data. It is unknown if current bioinformatic pipelines developed for analyses of target enrichment data can reliably exclude paralogous gene copies in data derived from the Angiosperm 353 probe kit.

Orthologs are genes related by descent from a common ancestor (due to a speciation event) and their evolutionary history tracks the phylogeny of species, while paralogs are products of gene duplication events. Theoretically, comparisons of paralogous copies of genes among species compromise phylogenetic inferences because the gene trees do not track speciation events, and hence, do not depict the true species relationships (Altenhoff et al. 2019; Fig. 1a). In the Hyb-Seq data or target enrichment data, in general, the paralogous genes might be "over-lumped" by assembly methods which use sequence similarity thresholds to define homology. The over-lumping of paralogs leads to inflation of sequence variation at those loci which may or may not

affect the inference of species relationships, but is expected to result in misestimation of branch lengths (and thus misestimates of divergence times). Therefore, excluding paralogs in phylogenetic studies using this type of data is pivotal, although paralogous gene sequences have value in other areas of comparative genomics (Madlung 2013; Limborg et al. 2016; McKinney et al. 2017). However, in Hyb-Seq data, orthologs and paralogs are often difficult to distinguish due to their high similarity in sequence identity (Altenhoff et al. 2019). All current pipelines for target enrichment data, including HybPiper (Johnson et al. 2016), PHYLUCE (Faircloth 2016), and SECAPR (Andermann et al. 2018), merely consider the sequence similarity in detecting paralogs.

A sequence similarity-based approach for calling paralogous genes may be sufficient for phylogenetic studies using custom designed probes based on orthologous sequences encompassing a closely related study group. However, for studies leveraging probes built from evolutionarily distant taxa from the focal group of investigation, especially in groups where gene and genome duplication are thought to be common such as plants, sequence similarity between contig and target genes alone may not be sufficient for removing all paralogs. Additional analyses of the sequence data may be needed to remove the potential paralogous sequences before performing phylogenetic analyses. In this study we propose supplementary criteria to sequence similarity for detecting and removing problematic paralogous gene data from Hyb-Seq by examining heterozygous sites within and among individuals in the aligned sequences. Low rates of shared heterozygous sites across all samples in a species-level dataset is expected under the assumption that polymorphisms among species are more likely to be fixed differences between paralogs over deep divergences (Eaton 2014, Eaton and Overcast 2020). Even at the shallow level of phylogenetic divergence (e.g. population genetics), high shared heterozygosity

109 across all samples within a locus may also be attributed to paralogs (Hohenlohe et al. 2011;
110 Harvey et al. 2016; McKinney et al. 2017). Additionally, a high number of heterozygous sites of
111 a locus within an individual may be considered an indicator of gene duplication events or
112 previously undetected polyploidy (Medina et al. 2019). Therefore, a high level of shared
113 heterozygosity at a site across individuals and high number of heterozygous sites within a locus
114 in an individual are both indicative of paralogy of the aligned gene sequences. In pipelines
115 developed for analyses of target enrichment data, usually an arbitrary cut-off of sequence identity
116 value between the contigs of a putative Hyb-Seq locus and the reference target gene is used to
117 determine if the locus contains paralogous sequences in an individual. Currently, HybPiper
118 (Johnson et al. 2016) uses BWA (Li and Durbin 2009) or BLASTx (Altschul et al. 1990) to
119 classify the raw reads into individual gene locus, followed by SPAdes (Bankevich et al. 2012) to
120 assemble the reads in a given individual into contigs (Fig. 1b). If multiple contigs with a >10x
121 coverage depth in an individual mapped to the same target gene with >85% sequence identity,
122 this target gene is marked for presence of paralogs in the individual (Fig. 1b), which can be
123 eliminated or addressed separately by investigators to determine its orthology to sequences of the
124 same locus of other individuals in subsequent analyses.

125 In addition, most pipelines for enrichment data implementing popular assemblers such as
126 SPAdes (Bankevich et al. 2012) and Abyss (Simpson et al. 2009) for sequence assembly can only
127 construct a single consensus sequence for a given locus in each individual (multiploidy) that
128 represents the most frequent base of each site among read variants. This approach loses all
129 information from heterozygous sites for identification of potential paralogs, which may result in
130 data containing phylogenetic noises from paralogous genes that can mislead the inference of
131 species relationship. Although SECAPR and scripts from Kates et al. (2018) can perform allele

phasing, all of the presently widely used pipelines for handling target enrichment data do not make use of the information from heterozygous sites to detect paralogous sequences. To make use of heterozygous sites, such as to detect paralogs or for phylogenetic inference, modification of existing pipelines for enrichment data is needed. In this study, we developed a new pipeline that generates degenerate coded sequences (retaining information of heterozygous sites) from Hyb-seq reads and uses criteria from both sequence similarity and quantity and distribution pattern of heterozygous sites for detection and cleaning of putative paralogs for downstream enrichment data analyses, which we call the Putative Paralogs Detection (PPD) pipeline (available on Github: https://github.com/Bean061/putative_paralog). We developed PPD by modifying HybPiper (see Methods) to code heterozygous sites in assemblies with IUPAC ambiguity codes, and to leverage these heterozygous sites for further filtering of putative paralogs (see details in Methods). In order to demonstrate the value of the new pipeline, we compared the number of putative paralogous loci detected by PPD and HybPiper and evaluated the influence of paralogs on phylogenetic and divergence time dating analyses using Hyb-Seq data from the Angiosperm 353 kit we generated for two diploid genera: *Castanea* (Chestnuts of Fagaceae) and *Hamamelis* (Witch-hazel of Hamamelidaceae). We further validated the paralogy of putative paralogous loci identified by PPD using a genome reference available for *Castanea*.

The chestnut genus *Castanea* Miller (Fagaceae) includes seven tree species, each restricted to eastern Asia (EA), eastern North America (ENA), or Europe. The species were divided into three sections (Dode 1908): section *Eucastanon* Dode, including the five species with three nuts per cupule: *C. mollissima* Blume and *C. seguinii* Dode from China and *C. crenata* Siebold & Zucc. from Japan, *C. dentata* (Marshall) Brokh. from North American, and *C. sativa* Mill. from Europe. Sections *Balanocastanon* Dode and *Hypocastanon* Dode each is monotypic

including a single species and both make fruits containing one nut per cupule. Section *Balanocastanon* contains *C. pumila* (L.) Mill. from North America and Section *Hypocastanon* contains *C. henryi* (Skan) Rehder & Wilson from China. Within *C. pumila*, two varieties were recognized by Johnson (1988) and Nixon (1997), *C. pumila* var. *pumila* in the southeastern United States and *C. pumila* var. *ozarkensis* (Ashe) A.E. Murray limited to the Ozark mountains. Phylogenetic studies of *Castanea* were previously conducted using data from six chloroplast regions in Lang et al. (2006, 2007). The studies found that sect. *Eucastanon* is paraphyletic. The witch-hazel genus *Hamamelis* L. (Hamamelidaceae) is also a small woody genus consisting of six species of shrubs and small trees, isolated in EA and ENA. The EA species include *H. mollis* Oliv. from eastern and southern China (Chang 1979; Zhang and Lu 1995) and *H. japonica* Siebold & Zucc. from Japan (Sargent 1890; Ohwi 1978). The ENA species include *H. virginiana* L., that is widely distributed from Canada to the Gulf coast (Bradford and Marsh 1977), *H. vernalis* Sarg., a species endemic to the Ozark Mountains in Arkansas, Missouri, and eastern Oklahoma (Bradford and Marsh 1977), *H. ovalis* S.W. Leonard that is restricted to a small area of Mississippi (Leonard 2006), and *H. mexicana* Standl. endemic to northeastern Mexico (Standley 1937), which is also known as *Hamamelis virginiana* var. *mexicana* (Standl.) C.Lane. A few phylogenetic studies of *Hamamelis* were previously conducted using data from ITS, ETS, waxy gene, and several plastid genes (Wen and Shi 1999; Li et al. 2000; Xie et al. 2010). However, the species relationships within *Hamamelis* have remained uncertain due to low nodal support values and short internal branches, especially regarding the relationships within the ENA clade. Therefore, results from the study also allow us to evaluate the previous phylogenetic hypotheses and further resolve the species relationships within these two genera.

MATERIALS & METHODS

Data Generation

Preparation of DNA samples.—We generated data from 15 samples of *Castanea*, seven samples of *Hamamelis*, and three samples of outgroups (Supplementary Table S1, available on Dryad), which covers all species of the two genera. Outgroup species were chosen based on their phylogenetic positions in Fagaceae and Hamamelidaceae, respectively inferred by Lang et al. (2006) and Xie et al. (2010). *Fothergilla* and *Parrotiopsis* were used as the outgroups of *Hamamelis* while *Quercus* was used as the outgroup of *Castanea*. Leaf samples were collected from the field or plants grown in arboreta or botanical gardens (Supplementary Table S1, available on Dryad). Fresh leaves were stored in silica gel to dry. The dry leaves were stored at -20 °C until they were used for the DNA extraction.

Total genomic DNAs were extracted from leaf samples using the CTAB protocol (Doyle 1991) with modification described in Cullings (1992) and Xiang et al. (1998). For leaf samples of *Castanea* that are rich in secondary compounds, they were washed five times with 0.8 mL of a washing buffer containing 10% polyethylene glycol, 0.35 M sorbitol, 50 mM Tris-HCl, 0.1% bovine serum albumin and 0.1% β-mercaptoethanol (Sakaguchi et al. 2018; Zhou et al. 2020) prior to DNA extraction with the modified CTAB method. The quality and quantity of DNA samples were first checked by 1% agarose gel electrophoresis followed by measurement on a Nanodrop spectrophotometer (ThermoFisher) and with a PicoGreen fluorescent dye assay (Life Technologies, ThermoFisher).

Library preparation of Angiosperm 353 gene enrichment and sequencing.— A total of 1000 ng DNA of each sample concentrated to ~35 μL was delivered to Rapid Genomics Lab (Gainesville,

Florida, USA) for Hyb-Seq library reconstruction and sequencing. The DNA samples were pooled for hybridization to biotinylated probes using the Angiosperm-353 v. 1 target capture kit (Johnson et al. 2019) available from Arbor Biosciences (Arbor Biosciences, Ann Arbor, Michigan, USA). Sequencing of DNAs pulled from the hybridization experiment was performed with Illumina MiSeq (Illumina, San-Diego, California. USA) to produce 2 x 150 bp paired end reads, as described in Gaynor et al. (2020).

Locus Data Assembly and MSA Generation

All samples were demultiplexed using Illumina's BCLtofastq by Arbor Biosciences. Raw sequencing reads were then cleaned and trimmed by Trimmomatic v.0.38 (Bolger et al. 2014) using parameters MAXINFO:100:0.5 and TRAILING:20. Subsequently, the HybPiper pipeline v. 3 (Johnson et al. 2016) was used to recover both coding sequences (CDS) and their flanking intron/non-coding regions. The process includes three major steps: using the nuclear sequences of Angiosperm 353 genes (Johnson et al. 2019) as the references to capture all the reads from sequenced accessions via the BWA option with default seed length k=19 (Li and Durbin 2009), applying the SPAdes (Bankevich et al. 2012) to assemble reads into long contigs, and implementing the intronerate.py module to recover "intron" and "supercontig" (CDS + intron fragments) sequences. Then, we used our PPD to generate multiple matrices to compare with those generated from HybPiper (see details below). To assess the phylogenetic and divergence time dating effects of paralogous genes we generated matrices consisting of supercontig sequences of three gene groups trimmed with PPD: orthologous loci, paralogous loci, or all loci. The supercontig matrices contained sequences of both coding and their flanking regions of the three respective gene groups. The original supercontig matrices derived from HybPiper were

retained for comparison. To build the matrices of orthologous genes, the paralogs called from HybPiper and PPD were manually removed from each all-gene matrix, while the matrices of paralogous genes included the paralogs detected by HybPiper and paralogs detected by PPD. Specifically, for the genes with paralog warning from HybPiper, we considered only those loci with warnings for at least two individuals as paralogs. This conservative approach followed Murphy et al. (2020), and was based on 1) the fact that the reference sequences for the Angiosperm 353 kit were putative single copy genes from diverse, evolutionarily distant taxa, and 2) the observation of a dissimilar sequence in one individual alone could be a random event or due to errors in sequencing or sequence assembly in that individual, rather than true paralogy. To allow different comparisons between the "*consensus*" matrices from HybPiper and "*degenerated*" matrices from PPD, we generated "*consensus*" matrices without (default of HybPiper), with gappy trimming (s6 of part 2 of PPD), and with all PPD trimming steps (all steps of part 2 of PPD, see details below). All data matrices and the relevant information are listed in Table 1.

Pipeline Description

The putative paralogs pipeline (PPD) includes two major parts: first, generating "*degenerated*" matrices, and second, trimming highly heterozygous sites, misaligned regions, and particularly gappy columns and detection of paralogous genes (Fig. 2).

In the first half of the PPD pipeline, the "*degenerated*" sequences are built for HybPiper derived supercontig or exon sequences (if the intron sequences were not captured or absent) of each locus using a bash script following Kates et al. (2018) (available on Github: https://github.com/Bean061/putative_paralogs). This involves using the "*consensus*" sequences

from HybPiper (Fig. 1b) as the references and mapping the raw reads back to the references in BWA with customized seed length according to the sequence length. As a higher seed length (BWA -k) value improves mapping quality (Robinson et al., 2017), we applied high seed length to ensure high quality mapping. Our sequencing method produced sequences of 150 bp for each read, we used a minimum seed length (-k) of 100 bp, instead of the default "-k" (19 bp). After mapping, the mapped duplicate reads are discarded using picard (https://broadinstitute.github.io/picard/). The program GATK (McKenna et al. 2010; DePristo et al. 2011) is then used to identify the variable sites using the HaplotypeCaller, with "-ploidy 2" parameter for diploid species, and SelectVariants functions. Finally, we use the FastaAlternateReferenceMaker function in GATK to convert the variable sites into the IUPAC coding to produce the "*degenerated*" (IUPAC) sequences for each gene.

The second half of the PPD pipeline trims alignments and detects paralogs, and includes 8 steps: s1) Resort gene files: Use all "*degenerated*" sequence files from every individual as the input, and then sort the degenerated sequences orthologous to the 353 reference genes in each sample into individual locus files according to gene names. s2) Sequence filtering: Filter the sequences with more than 5% (default) heterozygosity according to the percentage information of heterozygous sites in every sequence because a sequence with a high percentage of heterozygous sites may indicate sequencing or assembly errors of the particular locus. This setting can be changed by users with "-he" parameter. s3-s5) MSA generating: To obtain a better alignment result, the reference sequence of each locus is added for alignment using MAFFT (--adjustdirection –maxiterate 1000 --globalpair) (Katoh and Standley 2013). The reference sequences are removed before trimming of the aligned sequences in s6 and s7. s6) MSA trimming: Remove the gappy sites (i.e., sites missing in 50% or more individuals) using TrimAl

(default "-gt 0.51") (Capella-Gutierrez et al. 2009), a threshold based on the simulation study by Wiens and Morrill (2011) which showed that adding a set of characters with data for 50% of the species is either beneficial or harmless for phylogenetic study. We found the gappy regions were extensive and mostly at two ends spanning the intron/flanking regions of the gene sequence alignment in a locus, which might be attributed to erroneous assembly with a small number of raw reads in a few individuals. Therefore, we excluded these regions from the alignment to remove the influence of the gappy sites in phylogenetic analyses. The "-gt" parameter can also be customized (this parameter is identical to the "-gt" in TrimAl). s7) MSA further trimming: Detect and trim the hypervariable sites or regions using a sliding window method. The polymorphic sites in the ingroups meeting the requirement in each window were marked and then removed from all individuals (including the sites in outgroup species) by TrimAl. The maximum number of sites in a sliding window can be modified by the "-mi" parameter and sliding window length can be modified by the "-w" parameter in PPD. The default values for "-mi" and "-w" are 4 and 20, respectively, which represent if there are more than 4 polymorphic sites (not counting sites with heterozygous bases/degenerated sites) in a 20 bp sliding window (representing >25% variable sites) all of the polymorphic sites will be marked and removed by TrimAl. For polymorphic sites attributed solely to differences in sequences of the outgroups and meeting the requirement of more than 8 polymorphic sites (changeable via "-mo" parameter in PPD, default is 8) present in each 20 bp window, they are marked and replaced by a dash "-" in the sequence of the outgroup and the sites are not removed from any individuals to retain information likely phylogenetically informative among ingroup taxa. These criteria should be adjusted according to observation of the non-trimmed taxa MSA. We used the >25% cutoff for our data based on the assumption that such high rates of sequence variation in the 353 genes and

their flanking regions among our study ingroup species is unlikely true and may represent alignment ambiguity due to errors from sequence assembly. Our visual inspection of the BWA mapping result found the hypervariable sites had extremely low mapping quality, e.g., low depth of mapped reads (less than 5 reads) and many wrongly mapped reads. Including these sites would inflate sequence variation, thus, the branch length in phylogenetic inferences. s8) Paralog identification: Consider a locus as a paralog if it contains one or more heterozygous site(s) that are shared by 50% (default) or more individuals. The threshold of shared percentage and the number of heterozygous sites can be adjusted by the user using the "-hs" parameter and "-nh" parameter, respectively. For example, in Figure 2, a hypothetical MSA of a locus/gene (on the left side) shows sequence with high heterozygous sites (Sp1), a polymorphic site that is heterozygous in >50% samples/individuals of a diploid organism (labeled as polymorphic site 2), and a sequence containing a region with apparent alignment ambiguity due to error in contig assembly (shown as hypervariable sites compared to the rest). Identical heterozygous site(s) shared by over 50% individuals (Polymorphic site 2) in the MSA is used as the indication of presence of paralogs in the locus and is the criterion for calling putative paralogs in the PPD.

Phylogenetic Analyses

Concatenation-based tree.—Phylogenetic analyses of the concatenated Hyb-Seq data were performed for the supercontig data matrices of the three gene groups generated from PPD as well as supercontig data matrices of the orthologs derived from HybPiper listed in Table 1 using a maximum likelihood method implemented in IQ-TREE v. 1.6.12 (Nguyen et al. 2015) partitioned by genes. All analyses used the TESTNEW option to obtain the best molecular model per partition. UF bootstrap was applied to evaluate the topology (Hoang et al. 2018).To test the

316 congruence among different partition methods and phylogenetic methods, we also ran a
317 phylogenetic analysis with the best merged partitions suggested by ModelFinder using MFP-
318 MERGE in IQ-TREE (Lanfear et al. 2012) and conducted analyses without any partition using
319 RAxML (Stamatakis 2014) and MrBayes (Ronquist and Huelsenbeck 2003) for the degenerated
320 orthologous data matrices derived from PPD pipeline (for details, see Supplementary
321 Information, available on Dryad). The RAxML and MrBayes analyses above were all conducted
322 on the CIPRES Science Gateway Portal (Miller et al. 2010).
323
324 *Coalescent-based species tree.*—We used both ASTRAL-III (Zhang et al. 2018) and
325 SVDQuartets (Chifman and Kubatko 2014) to generate the coalescent-based species trees. For
326 the analyses with ASTRAL-III, we used gene trees from IQ-TREE for both genera as the inputs
327 and ran ASTRAL-III with the default parameters. For the analyses with SVDQuartets, we used
328 concatenated multilocus data as the input. Then, PAUP* v4.0a166 (Swofford 2003) was used to
329 generate a total of 100,000 quartets with 100 bootstrap replicates and then the quartet assembly
330 method QFM was used to produce a summary tree (Reaz et al. 2014), following Zhou et al.
331 (2020).
332 All concatenation-based trees and coalescent-based species trees were visualized and
333 edited in FigTree v.1.4.4 (Rambaut 2012) and edited with ggtree [R] (Yu et al. 2018) and Adobe
334 Illustrator 2020 (Adobe Systems, San Francisco, CA, USA).
335
336 *Divergence Time Analyses*
337 We employed BEAST2 2.6.2 (Bouckaert et al. 2014) to estimate the divergence times of
338 lineages within each genus. BEAST2 can consider information at heterozygous sites in

339 divergence time estimation. The divergence time analyses were conducted for orthologous and
340 paralogous matrices generated from the PPD (marked with asterisk in Table 1) to allow
341 comparisons and assess the effect of paralogous genes. The divergence time analysis was also
342 conducted for "*consensus*" supercontig matrices of orthologs with and without PPD trimming to
343 allow comparison between degenerated orthologous gene data derived from PPD and the
344 consensus orthologous gene data from the HybPiper alone. The stem ages of *Castanea* and
345 *Hamamelis* were constrained based on fossil evidence, as 66 to 72 Ma (lognormal) and 50 to 56
346 Ma (lognormal), respectively (for details, see Supplementary Information, available on Dryad).
347 Divergence time analyses were run under the GTR+Γ molecular model for all orthologous gene
348 matrices and HKY+Γ molecular model for paralogous gene matrices for both *Castanea* and
349 *Hamamelis*, the best models for each on the BIC values from jModelTest (Darriba et al. 2012).
350 An uncorrelated lognormal relaxed clock (Drummond et al., 2006) and the birth–death process
351 model (Stadler, 2010) were implemented in the analyses. To account for the fact that our
352 sampling in *Castanea* contained two samples per species, which violates assumptions of the BD
353 model, we performed an additional analysis of the orthologous gene data by using a single
354 sample per species to evaluate the impact of this violation. To facilitate comparisons among
355 datasets and between undated and dated phylogenies, we included the original sampling of
356 *Castanea* in divergence time analyses of all datasets. We run our analyses as a single
357 concatenated supermatrix, as divergence time analyses using concatenated unpartitioned
358 supermatrices compared with gene partitioned matrices of genomic data results in similar
359 divergence times, but the concatenated data sets were more efficient than the partitioned datasets
360 in attaining suitable effective sample sizes (Voloch and Schrago 2012). We set the mean
361 GrowthRate (net diversification rate) to have a uniform distribution with a range of 0–100, with

an initial value of 0.0, and the relative Death Rate (extinction rate/speciation rate) to have a 0–1 range, with an initial value of 0.5. These values were chosen based on the estimated average net diversification rate and extinction rate in plants (De Vos et al., 2015). Because constraints on node times can interact with constraints on other nodes and can also impact the divergence times of nodes that are elsewhere on the tree, we ran "empty" Markov Chain Monte Carlo analyses by adopting the prior settings but without using the sequence data to determine if the marginal densities of calibrated nodes matched the calibration densities, a desired property of a calibrated tree prior (Heled and Drummond 2012). These analyses yielded approximations to the prior distributions. To ensure that the prior distributions were well approximated, these "empty" MCMC runs all had effective sample sizes that exceeded 200. We found congruence between the priors and their approximations. Then, we ran the analyses with data for 200 million generations, with sampling of trees every 10,000 generations. Quality of the runs and parameter convergence were assessed using Tracer v.1.6.0 (Rambaut et al. 2018). The maximum credibility tree of median heights was then constructed using TreeAnnotator after discarding 20% trees as burn-in.

Assessment of PPD Success Rate on Paralogs Identification

To test whether the putative paralogs detected by PPD were true paralogs and assess the false positive and false negative rates of PPD in identifying paralogs, we conducted nucleotide BLAST (Altschul et al. 1990) search to determine if the putative orthologs and paralogs would map to one or more regions of reference genome sequences. One *Castanea* species has a published genome (*C. mollissima* ASM1418300v1 from Wang et al. 2020) but no species of *Hamamelis* has genome sequences available. Because considering BLAST results using distant genome references may not reflect gene paralogy correctly, we assessed the success rate of PPD

in identification of paralogous genes only in *Castanea* samples using the *Castanea* reference genome. We considered a locus to be confirmed as paralogous when its sequence from any *Castanea* sample had two or more BLAST hits on the reference genome and/or had a BLAST hit to a genome location different from that of other samples with 90 percent of identity with at least 500 bp mapping length in the separate regions of the reference genome. We calculated the success rate of PPD in paralogous gene identification as the number paralogous loci confirmed by the BLAST mapping analyses divided by the total number of paralogous loci identified by PPD. We also assessed the failure rate of PPD in calling paralogous genes by mapping the pooled sequences of all species of a putative orthologous locus to the reference genome. If sequences of a gene locus are mapped to more than one region in the reference genome, we recorded it as a case of false orthology. We also evaluated if false orthology and false paralogs influenced our phylogenetic analyses by repeating IQ-TREE analyses described above on a matrix that contained PPD orthologs and putative false paralogs but excluding any putative false orthology from BLAST results.

RESULTS

The number of loci, alignment length, average length per locus, total hypervariable sites removed, number of segregating sites, and number of parsimony informative sites varied among the three gene groups and between genera (Table 1). We found no sequences with excess heterozygosity and thus no sequences were removed from our data due to the presence of excess within-individual heterozygosity (5% or more). In the consensus matrices generated from HybPiper, approximately an average of 1120 bp in *Castanea* and 446 bp in *Hamamelis* were removed from each locus through the gap-trimming step in PPD. Through the PPD sliding

408 window trimming process, approximately an average of 163 bp and 110 bp hypervariable sites
409 from each locus were detected and removed from *Castanea* and *Hamamelis*, respectively.
410
411 *Paralog Detection in Hyb-Seq Data*
412 The gene matrices generated by HybPiper had paralog warning for 11 loci shared in two
413 or more individuals (Gene 6048, 6954, 4951, 4724, 5940, 6387, 6570, 7583, 7324, 5138, 5941)
414 out of a total of 344 genes sequenced in *Castanea*, but only two putative paralogs (Gene 5463,
415 5347) out of 346 genes in *Hamamelis* were identified based on the same criteria. In contrast, our
416 PPD pipeline (in conjunction with HybPiper) detected 48 and 27 paralogs in *Castanea and*
417 *Hamamelis*, respectively (Table 1). We found 31 (77.5%) out of 40 paralogs from PPD had
418 multiple hits to the *Castanea* reference genome, while 9 (22.5%) paralogs had one hit based on
419 the BLAST results (Table 2; Supplementary Tables S2 and S3, available on Dryad). In
420 orthologous genes detected by PPD, we found 255 (83.9%) out of 304 orthologs had only a
421 single hit (i.e. all samples mapped only to a single region of the genome), while 46 (15.1%)
422 putative orthologs had multiple hits (Table 2; Supplementary Tables S4 and S5, available on
423 Dryad). Phylogenetic analyses that also excluded orthologs with multiple BLAST hits and
424 included paralogs with single BLAST hits were qualitatively the same as all other PPD analyses
425 described below (Supplementary Fig. S1, available on Dryad). As a comparison, we found 11
426 paralogs by HybPiper, eight of which differed from paralogs from PPD. Four (36.4%) out of 11
427 paralogs from HybPiper had multiple hits to the *Castanea* reference genome, while six (54.5%)
428 paralogs had one hit and one putative paralog had no hits (Table 2). Among the 333 orthologous
429 genes from HybPiper, 258 (77.5%) had single hit, 73 (21.9%) had multiple hits, and 2 (0.6%)
430 had no hits (Table 2).

Phylogenetic Analyses of Orthologous Gene Data

The phylogenetic analyses of the orthologous gene data from PPD using IQ-TREE (with gene partition and best merged partition), RAxML, and MrBayes resulted in the same tree topologies with strong nodal support in both *Castanea* (Fig. 3a; Supplementary Figs. S2- S4 available in Dryad) and *Hamamelis* (Fig. 3b and Supplementary Figs. S5 - S7, available on Dryad). The coalescent-based species trees reconstructed from ASTRAL and SVDQuartets for each genus also had the same topology identical to the concatenation-based tree (Fig. 4). In *Castanea*, the reciprocal monophyly of species from EA and ENA were recovered for each region, and the European species *C. sativa* was placed as the sister to the American clade (Fig. 4a). In *Hamamelis*, species from ENA form a monophyletic group sister to *H. japonica* with *H. mollis* diverging out first, sister to the remaining species. However, the node connecting the ENA clade and *H. japonica* was not well supported in ASTRAL (0.59) but well supported in SVDQuartets (90) (Fig. 4b).

Phylogenetic analysis of the orthologous gene data from HybPiper alone with and without PPD trimming steps resulted in different results in the two genera considered. In *Castanea*, the same topology was recovered from orthologous gene data for HybPiper matrices with and without PPD trimming, and this topology was the same as the topology recovered from the full PPD pipeline (Compare Figs. 3a, 3c, and 3e). In *Hamamelis*, the analysis of the untrimmed matrix resulted in a tree with a topology different from the tree from the PPD and trimmed HybPiper data (Compare Figs. 3b, 3d, and 3f). In both genera, the branch lengths in HybPiper data-based trees were substantially longer than trees based on the PPD data, especially in the trees from the untrimmed HybPiper consensus data.

Divergence Time Analyses of Orthologous Genes

Castanea.—Divergence time analyses of the PPD-derived data including all samples (i.e., degenerated supercontigs of orthologous genes) estimated the crown age of the genus (splitting of the EA and ENA clades) as the early Miocene (17.9 Ma, 95% HPD: 14.3-21.8 Ma). Within the genus, other divergence occurred in the mid-Miocene and late Miocene (Fig. 5a and Supplementary Table S6, available on Dryad). The European chestnut (*C. sativa*) diverged from the two ENA species in the mid-Miocene (13.6 Ma, 95% HPD: 10.9-16.7 Ma). The divergence times estimated from analysis with one sample per species were highly similar to those based on full sampling (two samples per species) for *Castanea*, with differences of median values < 1 million years (Supplementary Fig. S8, available on Dryad).

Divergence times (median values) estimated from the HybPiper-derived data were approximately 11 million years (untrimmed) and two million years (trimmed) older, respectively, for all the nodes (Figs. 5a, 5c, and 5g and Supplementary Table S6, available on Dryad). The divergence times estimated from the paralogous genes were two to a few million years older than the estimates based on the orthologous gene data (Fig. 5e, and Supplementary Table S6, available on Dryad).

Hamamelis.—Divergence time analyses of the PPD-derived data showed the crown node of *Hamamelis* (splitting of *H. mollis* from the remaining species) was dated back to the late Oligocene (e.g., 27.6 Ma with the 95% HPD as 24.0-31.6 Ma; Fig. 5b and Supplementary Table S6, available on Dryad). The divergence of *H. japonica* from the ENA clade was dated to the early Miocene (e.g., 23.3 Ma, with the 95% HPD as 20.2-26.7 Ma; Fig. 5b and Supplementary Table S6, available on Dryad). Divergence events within the American clade were dated to the

late Miocene for *H. virginiana* and the Pliocene for the other species (Fig. 5b and Supplementary Table S6, available on Dryad). Similarly, the divergence times estimated from HybPiper-derived data were approximately 6 to 10 million years (untrimmed) and up to three million years (trimmed) older, respectively, for all nodes (Fig. 5d and 5h; Supplementary Table S6, available on Dryad).

Divergence time analyses of the paralogous gene data detected by PPD showed the median were highly similar at some nodes but younger or older at other nodes with differences within four or five million years, compared to the estimates from the "*degenerated*" orthologous gene data (Fig. 5f and Supplementary Table S6, available on Dryad). However, the 95% HPD were much higher at all nodes, indicating greater uncertainty.

DISCUSSION

Impacts of Paralogs and the Value of the PPD

Our results showed that our new pipeline (PPD) identified many more putative paralogs than HybPiper. Although the "*consensus*" sequence data generated from HybPiper may produce the phylogenetic tree with the same topology as the tree from the "*degenerated*" sequence data derived from PPD, the HybPiper data contained many more "false" phylogenetic informative sites (due to the presence of paralogous genes and consensus coding of the sequences), resulting in longer branches affecting divergence time estimation (Figs. 3 and 5; Supplementary Table S6, available on Dryad). The sequence data with better cleaning of paralogs and coded with the "*degenerated*" method are advantageous for phylogenomic studies, as they contain more accurate information for phylogenetic and divergence time estimations. Comparisons of the PPD data with "*consensus*" data with and without PPD trimming steps (Figs. 3 and 5) indicated that

the observed differences in branch lengths and divergence times cannot be explained by differences in trimming alone and sequence coding and paralogs also affected branch lengths and divergence time estimation. Furthermore, our phylogenetic analyses of the loci containing paralogous genes often resulted in phylogenies different from those inferred from data of the orthologous genes in *Hamamelis* (Supplementary Fig. S9, available on Dryad). The divergence times estimated from data including potential paralogous genes (i.e., the "*consensus*" data matrices from HybPiper) or from the paralogous genes identified from PPD are older and have larger HPDs, likely due to the additional variable sites introduced by gene paralogy (Fig. 5 and Supplementary Table S6, available on Dryad). Our results clearly highlighted the negative impacts of paralogous gene content in phylogenetic analyses and that paralogous gene content either inflates estimates of divergence time or increases uncertainty of divergence time estimation in *Castanea* and *Hamamelis* (Fig. 5 and Supplementary Table S6, available on Dryad). Comparisons of the PPD data with the "*consensus*" and "*untrimmed*" consensus data from HybPiper further indicated that the effects of sequence trimming on branch lengths and divergence time estimation were major, greater than the influences of sequence coding and paralogs in our case (Figs. 3 and 5). These results together strongly support that additional steps following HybPiper to "polish" data from Hyb-Seq of Angiosperm 353 probe kit are necessary before phylogenetic and downstream analyses. Moreover, we show that the PPD pipeline can effectively clean alignments with user-defined trimming and identify paralogs in these alignments to produce higher quality data for phylogenetic and divergence time dating analyses. The "*degenerated*" matrix generated from the PPD using the IUPAC ambiguity codes are suitable for a wide range of modern phylogenetic tools for phylogeny and divergence time estimation, including RAxML (Stamatakis 2014), IQ-TREE (Nguyen et al. 2015), SVDQuartets

(Chifman and Kubatko 2014), BEAST2 (Bouckaert et al. 2014) that has an option to treat ambiguity-coded positions as informative.

Accuracy Rate and Caveats of PPD in Paralogs Identification

Through BLAST mapping analysis with the *Castanea mollissima* genome, the paralogy of most of the PPD identified paralogous loci (31 out of 40 at a rate of 77.5%) were confirmed by two or more hits. The remaining nine paralogs each had a single hit in the genome (i.e. all samples mapped only to a single region), which represented false-positive paralogs, may be explained by loss of the duplicated paralogous loci in the reference genome and/or incompleteness of the reference genome. Additional *Castanea* genomes that may become available in the future will help further test this hypothesis. Alternatively, small-scale duplication events (e.g., Hudson et al. 2011; Carretero-Paulet and Fares 2012; Rensing 2014) that are prevalent in *Castanea* plants may be missed based on the settings we used for BLAST (such as a 500 bp length), leading to the false classification of a putative paralog as having only a single hit. We found that five out of these nine loci have only one heterozygous site shared by >50% individuals. The single shared heterozygous site in these five paralogs could be a result of occurrence by chance or sequencing errors. If users want to minimize such potentially false identification of paralogs and they can use a more conservative approach by increasing the number of heterozygous sites shared by >50% individuals. However, this may result in the potential of missing true paralogous loci. If no reference genome is available for verification of paralogy of loci, and given that sequences for numerous loci are available from Hyb-Seq for phylogenetic analyses, we recommend a more aggressive approach to removing paralogs, such as the one adopted in our study.

Our mapping analysis also indicated that PPD outperformed HybPiper alone at identifying true orthologs. We found 255 out 303 (83.9%) orthologous genes identified by PPD were true orthologs (evidenced by a single hit in the BLAST analysis), compared with only 77.5% from HybPiper alone. Additionally, 46 of the orthologous genes from PPD had two hits (15.1%), indicating paralogy of these loci according to our mapping criterion, while 73 (21.9%) of putative orthologs from HybPiper alone had multiple hits. This may indicate that both HybPiper and PPD do not remove all potential paralogs, but with only a single reference genome available, it is also possible these putative orthologs mapping multiple times could reflect errors in reference genome assemblies. Regardless of the origin of these putative paralogs missed by PPD, excluding them from phylogenetic analyses did not result in substantial differences in phylogenetic results between the original orthologous PPD matrix and one without these genes, indicating that a small percentage of "false" orthologs is tolerable. However, researchers may choose to validate the PPD identified paralogs and orthologs for their taxa with reference genome available and further refine the data, as done with *Castanea* in our study Overall, compared to HybPiper, PPD generated more accurate orthologous gene data for phylogenetic and downstream analyses (Table 1 and Supplementary Table S7, available on Dryad).

Taxonomy and Relationships within Castanea and Hamamelis

Our phylogenetic data do not agree with the morphology-based classification scheme of three sections in *Castanea* (Sect. *Eucastanon*, Sect. *Balanocastanon*, and Sect. *Hypocastanon*) (Dode 1908; Johnson 1988). Our result indicated that Sect. *Eucastanon* that included *C. dentata*, *C. sativa*, *C. mollissima*, *C. seguinii*, and *C. crenata* is paraphyletic and the character of one nut per cupule in *C. pumila* (ENA) and *C. henryi* (EA) is homoplasy. Our results also support that

569 ENA clade is sister to European *C. sativa* with high support value (Figs. 3a and 4a). The
570 taxonomic status of the Allegheny chinkapin (*C. pumila*) and the Ozark chinkapin has been
571 disputed (Johnson 1988; Nixon 1997). Johnson (1988) considered the Ozark chinkapin as a
572 variety of *C. pumila*, while Nixon (1997) regarded it as a separate species *C. ozarkensis*. In our
573 study, all individuals representing *C. pumila* including the Ozark chinkapins formed a
574 monophyletic group sister to *C. dentata* with strong support. Therefore, our phylogenomic study
575 does not support the recognition of *C. ozarkensis* as a distinct species. However, the hypothesis
576 should be further tested with population level sampling of related taxa.

577 In *Hamamelis*, our result suggested a similar topology with previous phylogenies using
578 data from ITS, ETS, waxy gene, and several plastid genes (Wen and Shi 1999; Li et al. 2000; Xie
579 et al. 2010), which showed *H. mollis* diverged first, followed by the divergence between *H.*
580 *japonica* and ENA clade. The ENA clade was a well-supported monophyletic clade. Different
581 from previous studies, our concatenation-based tree showed a well resolved relationship among
582 ENA clade using nuclear gene data, indicating *H. virginiana* is the first diverged species,
583 followed by the divergence of *H. vernalis,* and *H. mexicana* is sister to *H. ovalis* (Fig. 3b).
584 However, our coalescent-based species tree showed a different topology within the ENA clade,
585 uniting *H. ovalis* and *H. vernalis* as the sister group but with low support values (Fig. 4b). This
586 conflict suggests there might be incomplete lineage sorting or gene flow among these three taxa
587 in North America. The node connecting *H. japonica* and ENA clade is also relatively low in the
588 species tree reconstructed with ASTRAL (0.59; Fig. 4b), indicating another phylogenetic conflict
589 among gene trees and the possibility of ancient gene flow or incomplete lineage sorting.

590 In conclusion, PPD, the pipeline we have described here, improves the quality of data
591 obtained from Hyb-Seq for phylogenomic analyses through detection of additional paralogous

genes and removal of hypervariable regions. Through empirical studies in *Castanea* and *Hamamleis*, our study demonstrated that data derived from HybPiper without the filtering steps implemented in PPD biased phylogenetic and divergence time estimation. Although our results focused on expanding HybPiper to improve detection of paralogs, our study also highlights the importance of accounting for potential paralogous genes in phylogenomic studies. As such, we recommend that phylogenomic analyses account for paralogs, such as through our PPD tool, particularly when the study group of interest belongs to lineages where gene duplication could be a concern.

SUPPLEMENTARY MATERIAL

Data available from the Dryad Digital Repository:
https://doi.org/10.5061/dryad.ttdz08kwq

Demultiplexed sequence data are available for download from the NCBI Sequence Read Archive (SRA) (BioProject PRJNA670453).

FUNDING

The study was supported by an NSF grant of the United States DEB – 1442161 to QY(J) Xiang. This work was also benefited from the USDA National Institute of Food and Agriculture, Hatch project 02718. J. Soghigian was supported by NSF DEB – 1754376.

ACKNOWLEDGEMENTS

We thank the Soltis lab at Florida Museum of Natural History, CX Fu lab at Zhejiang University, Gao lab at Kunming Institute of Botany, Chinese Academy of Sciences, JC Raulston

615 Arboretum, Arnold Arboretum, University of Washington Botanic Gardens, P Jones at Sarah
616 Duke Garden, UNC herbarium, J Lee from Korea research institute of bioscience and
617 biotechnology and Y Ru from Heidelberg University for providing some leaf tissues, JL Thorne
618 from North Carolina State University and M Johnson from Texas Tech University for discussion
619 of the PPD pipeline. Special thanks to three anonymous reviewers and editors for precious
620 suggestions. We also thank SR Manchester and K Pigg for discussion of *Castanea* and
621 *Hamamelis* fossils. Finally, we thank Z Du from Wuhan Botanical Garden to run pilot tests for
622 the PPD with data from *Cornus* and *Aesculus*.

623

624 REFERENCES

625 Altenhoff A.M., Glover N.M., Dessimoz C. 2019. Inferring orthology and paralogy. In
626 Evolutionary Genomics. New York: Humana Press. p. 149–175.
627 Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. 1990. Basic local alignment search
628 tool. J. Mol. Biol. 215:403–410.
629 Andermann T., Cano Á., Zizka A., Bacon C., Antonelli A. 2018. SECAPR—a bioinformatics
630 pipeline for the rapid and user-friendly processing of targeted enriched Illumina sequences, from
631 raw reads to alignments. PeerJ. 6:e5175.
632 Banker S.E., Lemmon A.R., Hassinger A.B., Dye M., Holland S.D., Kortyna M.L., Ospina O.E.,
633 Ralicki H., Lemmon E.M. 2020. Hierarchical Hybrid Enrichment: Multitiered Genomic Data
634 Collection Across Evolutionary Scales, With Application to Chorus Frogs (*Pseudacris*). Syst.
635 Biol. 69:756–773.
636 Bankevich A., Nurk S., Antipov D., Gurevich A.A., Dvorkin M., Kulikov A.S., Lesin V.M.,
637 Nikolenko S.I., Pham S., Prjibelski A.D., Pyshkin A.V., Sirotkin A.V., Vyahhi N., Tesler G.,

638 Alekseyev M.A., Pevzner P.A. 2012. SPAdes: A New Genome Assembly Algorithm and Its
639 Applications to Single-Cell Sequencing. J. Comput. Biol. 19:455–477.
640 Barrow L.N., Lemmon A.R., Lemmon E.M. 2018. Targeted Sampling and Target Capture:
641 Assessing Phylogeographic Concordance with Genome-wide Data. Syst. Biol. 67:979–996.
642 Bolger A.M., Lohse M., Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence
643 data. Bioinformatics. 30:2114–2120.
644 Bouckaert R., Heled J., Kuhnert D., Vaughan T., Wu C.H., Xie D., Suchard M.A., Rambaut A.,
645 Drummond A.J. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS
646 Comput. Biol. 10:e1003537.
647 Bradford J.L., Marsh D.L. 1977. Comparative Studies of the Witch Hazels *Hamamelis virginiana*
648 and *H. vernalis*. J. Ark. Acad. Sci. 31(1): 29-31.
649 Capella-Gutierrez S., Silla-Martinez J.M., Gabaldon T. 2009. trimAl: a tool for automated
650 alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 25:1972–1973.
651 Carretero-Paulet L., Fares M.A. 2012. Evolutionary Dynamics and Functional Specialization of
652 Plant Paralogs Formed by Whole and Small-Scale Genome Duplications. Mol. Biol. Evol.
653 29:3541–3551.
654 Chang H.T. 1979. Hamamelidaceae. In: Florae Reipublicae Popularis Sinicae. Beijing: Science
655 Press. 35(2): 36-116.
656 Chifman J., Kubatko L. 2014. Quartet inference from SNP data under the coalescent model.
657 Bioinformatics. 30:3317–3324.
658 Cullings K.W. 1992. Design and testing of a plant-specific PCR primer for ecological and
659 evolutionary studies. Mol. Ecol. 1:233–240.

660 Darriba D., Taboada G.L., Doallo R., Posada D. 2012. jModelTest 2: more models, new
661 heuristics and parallel computing. Nat. Methods. 9:772.
662 DePristo M.A., Banks E., Poplin R., Garimella K.V., Maguire J.R., Hartl C., Philippakis A.A.,
663 del Angel G., Rivas M.A., Hanna M., McKenna A., Fennell T.J., Kernytsky A.M., Sivachenko
664 A.Y., Cibulskis K., Gabriel S.B., Altshuler D., Daly M.J. 2011. A framework for variation
665 discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43:491–498.
666 De Vos J.M., Joppa L.N., Gittleman J.L., Stephens P.R., Pimm S.L. 2015. Estimating the normal
667 background rate of species extinction. Conserv. Biol. 29:452–462.
668 Dode L.A. 1908. Notes dendrologiques. Paris: Au Siège de la Société.p. 1-166.
669 Dodsworth S., Pokorny L., Johnson M.G., Kim J.T., Maurin O., Wickett N.J., Forest F., Baker
670 W.J. 2019. Hyb-Seq for Flowering Plant Systematics. Trends Plant Sci. 24:887–891.
671 Dong Y., Chen S., Cheng S., Zhou W., Ma Q., Chen Z., Fu C.-X., Liu X., Zhao Y., Soltis P.S.,
672 Wong G.K.-S., Soltis D.E., Xiang Q.-Y. 2019. Natural selection and repeated patterns of
673 molecular evolution following allopatric divergence. eLife. 8:e45199.
674 Doyle J. 1991. DNA protocols for plants. Molecular techniques in taxonomy. Springer-Verlag. p.
675 283–293.
676 Du Z.-Y., Harris A., Xiang Q.-Y. (Jenny). 2020. Phylogenomics, co-evolution of ecological niche
677 and morphology, and historical biogeography of buckeyes, horsechestnuts, and their relatives
678 (Hippocastaneae, Sapindaceae) and the value of RAD-Seq for deep evolutionary inferences back
679 to the Late Cretaceous. Mol. Phylogenet. Evol. 145:106726.
680 Drummond A.J., Ho S.Y., Phillips M.J., Rambaut A. 2006. Relaxed phylogenetics and dating
681 with confidence. PLoS Biol. 4:e88.

682 Eaton D.A.R. 2014. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses.
683 Bioinformatics. 30:1844–1849.
684 Eaton D.A.R., Overcast I. 2020. ipyrad: Interactive assembly and analysis of RADseq datasets.
685 Bioinformatics. 36:2592–2594.
686 Faircloth B.C. 2016. PHYLUCE is a software package for the analysis of conserved genomic
687 loci. Bioinformatics. 32:786–788.
688 Faircloth B.C., McCormack J.E., Crawford N.G., Harvey M.G., Brumfield R.T., Glenn T.C.
689 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple
690 evolutionary timescales. Syst. Biol. 61:717–726.
691 Faircloth B.C., Sorenson L., Santini F., Alfaro M.E. 2013. A Phylogenomic Perspective on the
692 Radiation of Ray-Finned Fishes Based upon Targeted Sequencing of Ultraconserved Elements
693 (UCEs). PLoS ONE. 8:e65923.
694 Fu C.N., Mo Z.Q., Yang J.B., Ge X.J., Li D.Z., Xiang Q.J., Gao L.M. 2019. Plastid
695 phylogenomics and biogeographic analysis support a trans-Tethyan origin and rapid early
696 radiation of Cornales in the Mid-Cretaceous. Mol. Phylogenet. Evol. 140:106601.
697 Gaynor M.L., Fu C., Gao L., Lu L., Soltis D.E., Soltis P.S. 2020. Biogeography and ecological
698 niche evolution in Diapensiaceae inferred from phylogenetic analysis. J. Syst. Evol. 58(5): 646–
699 662.
700 Harvey M.G., Smith B.T., Glenn T.C., Faircloth B.C., Brumfield R.T. 2016. Sequence Capture
701 versus Restriction Site Associated DNA Sequencing for Shallow Systematics. Syst. Biol.
702 65:910–924.
703 Heled J., Drummond A.J. 2012. Calibrated Tree Priors for Relaxed Phylogenetics and
704 Divergence Time Estimation. Syst. Biol. 61:138–149.

Hoang D.T., Chernomor O., Von Haeseler A., Minh B.Q., Vinh L.S. 2018. UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35:518–522.

Hohenlohe P.A., Amish S.J., Catchen J.M., Allendorf F.W., Luikart G. 2011. Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout: SNP Dsicovery: Next Generation Sequencing. Mol. Ecol. Resour. 11:117–122.

Hudson C.M., Puckett E.E., Bekaert M., Pires J.C., Conant G.C. 2011. Selection for Higher Gene Copy Number after Different Types of Plant Gene Duplications. Genome Biol. Evol. 3:1369–1380.

Johnson G.P. 1988. Revision of *Castanea* sect *Balanocastanon* (Fagaceae). J. Arnold Arbor.:25–49.

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.

Johnson M.G., Pokorny L., Dodsworth S., Botigué L.R., Cowan R.S., Devault A., Eiserhardt W.L., Epitawalage N., Forest F., Kim J.T., Leebens-Mack J.H., Leitch I.J., Maurin O., Soltis D.E., Soltis P.S., Wong G.K., Baker W.J., Wickett N.J. 2019. A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering. Syst. Biol. 68:594–606.

Kates H.R., Johnson M.G., Gardner E.M., Zerega N.J.C., Wickett N.J. 2018. Allele phasing has minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in a case study of *Artocarpus*. Am. J. Bot. 105:404–416.

727 Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7:
728 improvements in performance and usability. Mol. Biol. Evol. 30:772–80.
729 Lanfear R., Calcott B., Ho S.Y.W., Guindon S. 2012. PartitionFinder: Combined Selection of
730 Partitioning Schemes and Substitution Models for Phylogenetic Analyses. Mol. Biol. Evol.
731 29:1695–1701.
732 Lang P., Dane F., Kubisiak T.L. 2006. Phylogeny of *Castanea* (Fagaceae) based on chloroplast
733 trnT-L-F sequence data. Tree Genet. Genomes. 2:132–139.
734 Lang P., Dane F., Kubisiak T.L., Huang H. 2007. Molecular evidence for an Asian origin and a
735 unique westward migration of species in the genus *Castanea* via Europe to North America. Mol.
736 Phylogenet. Evol. 43:49–59.
737 Larridon I., Villaverde T., Zuntini A.R., Pokorny L., Brewer G.E., Epitawalage N., Fairlie I.,
738 Hahn M., Kim J., Maguilla E., Maurin O., Xanthos M., Hipp A.L., Forest F., Baker W.J. 2020.
739 Tackling Rapid Radiations With Targeted Sequencing. Front. Plant Sci. 10:1655.
740 Leache A.D., Chavez A.S., Jones L.N., Grummer J.A., Gottscho A.D., Linkem C.W. 2015.
741 Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus
742 restriction site associated DNA sequencing. Genome Biol. Evol. 7:706–719.
743 Lemmon A.R., Emme S.A., Lemmon E.M. 2012. Anchored Hybrid Enrichment for Massively
744 High-Throughput Phylogenomics. Syst. Biol. 61:727–744.
745 Lemmon E.M., Lemmon A.R. 2013. High-Throughput Genomic Data in Systematics and
746 Phylogenetics. Annu. Rev. Ecol. Evol. Syst. 44:99–121.
747 Leonard S. 2006. A New Species of Witch-Hazel (*Hamamelis*: Hamamelidaceae) Apparently
748 Endemic to Southern Mississippi. SIDA, Contributions to Botany. 22(2):849–856.

749 Léveillé-Bourret É., Starr J.R., Ford B.A., Moriarty Lemmon E., Lemmon A.R. 2018. Resolving
750 Rapid Radiations within Angiosperm Families Using Anchored Phylogenomics. Syst. Biol.
751 67:94–112.
752 Li H., Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.
753 Bioinformatics. 25:1754–1760.
754 Li J., Bogle A.L., Klein A.S., Donoghue M.J. 2000. Phylogeny and biogeography of *Hamamelis*
755 (Hamamelidaceae). Harv. Pap. Bot. 5:171–178.
756 Limborg M.T., Seeb L.W., Seeb J.E. 2016. Sorting duplicated loci disentangles complexities of
757 polyploid genomes masked by genotyping by sequencing. Mol. Ecol. 25:2117–2129.
758 Madlung A. 2013. Polyploidy and its effect on evolutionary success: old questions revisited with
759 new tools. Heredity. 110:99–104.
760 McCartney D.L., Walker R.M., Morris S.W., McIntosh A.M., Porteous D.J., Evans K.L. 2016.
761 Identification of polymorphic and off-target probe binding sites on the Illumina Infinium
762 MethylationEPIC BeadChip. Genom. Data. 9:22–24.
763 McCormack J.E., Hird S.M., Zellmer A.J., Carstens B.C., Brumfield R.T. 2013. Applications of
764 next-generation sequencing to phylogeography and phylogenetics. Mol. Phylogenet. Evol.
765 66:526–538.
766 McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K.,
767 Altshuler D., Gabriel S., Daly M., DePristo M.A. 2010. The Genome Analysis Toolkit: A
768 MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res.
769 20:1297–1303.

McKinney G.J., Waples R.K., Seeb L.W., Seeb J.E. 2017. Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. Mol. Ecol. Resour. 17:656–669.

Medina R., Johnson M.G., Liu Y., Wickett N.J., Shaw A.J., Goffinet B. 2019. Phylogenomic delineation of Physcomitrium (Bryophyta: Funariaceae) based on targeted sequencing of nuclear exons and their flanking regions rejects the retention of Physcomitrella, Physcomitridium and Aphanorrhegma. J. Syst. Evol. 57:404–417.

Miller M. A., Pfeiffer W., Schwartz T., 2010, Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In *Gateway Computing Environments Workshop (GCE)*. IEEE. p. 1–8. DOI: 10.1109/GCE.2010.5676129.

Murphy B., Forest F., Barraclough T., Rosindell J., Bellot S., Cowan R., Golos M., Jebb M., Cheek M. 2020. A phylogenomic analysis of *Nepenthes* (Nepenthaceae). Mol. Phylogenet. Evol. 144:106668.

Nguyen L.T., Schmidt H.A., Von Haeseler A., Minh B.Q. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32:268–274.

Nixon K. 1997. *Castanea*. In Flora of North America North of Mexico. New York: Oxford University Press. 3:439–442.

Ohwi J. 1978. *Hamamelis*. In: Flora of Japan. Tokyo: Shibundo Co. Ltd. Publishers. p. 1–724.

One Thousand Plant Transcriptomes Initiative. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. Nature. 574:679–685.

Pais A.L., Li X., Jenny Xiang Q.-Y. 2018. Discovering variation of secondary metabolite diversity and its relationship with disease resistance in *Cornus florida* L. Ecol. Evol. 8:5619–5636.

Pais A.L., Whetten R.W., Xiang Q.-Y.J. 2017. Ecological genomics of local adaptation in *Cornus florida* L. by genotyping by sequencing. Ecol. Evol. 7:441–465.

Rambaut A. 2012. FigTree v1. 4. Available: https://github.com/rambaut/figtree/releases/tag/v1.4.4.

Rambaut A., Drummond A.J., Xie D., Baele G., Suchard M.A. 2018. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst. Biol. 67:901–904.

Reaz R., Bayzid M.S., Rahman M.S. 2014. Accurate phylogenetic tree reconstruction from quartets: a heuristic approach. PLoS One. 9:e104008.

Rensing S.A. 2014. Gene duplication as a driver of plant morphogenetic evolution. Current Opinion in Plant Biology. 17:43–48.

Robinson K.M., Hawkins A.S., Santana-Cruz I., Adkins R.S., Shetty A.C., Nagaraj S., Sadzewicz L., Tallon L.J., Rasko D.A., Fraser C.M., Mahurkar A., Silva J.C., Dunning Hotopp J.C. 2017. Aligner optimization increases accuracy and decreases compute times in multi-species sequence data. Microb. Genom. 3:e000122.

Ronquist F., Huelsenbeck J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.

Sakaguchi S., Takahashi D., Setoguchi H., Isagi Y. 2018. Genetic structure of the clonal herb Tanakaea radicans (Saxifragaceae) at multiple spatial scales, revealed by nuclear and mitochondrial microsatellite markers. Plant Species Biol. 33:81–87.

Sargent C.S. 1890. Hamamelidaceae-Sapotaceae. In The sylva of North America. New York: Peter Smith.

Shee Z.Q., Frodin D.G., Cámara-Leret R., Pokorny L. 2020. Reconstructing the Complex Evolutionary History of the Papuasian Schefflera Radiation Through Herbariomics. Front. Plant Sci. 11:258.

Simpson J.T., Wong K., Jackman S.D., Schein J.E., Jones S.J.M., Birol I. 2009. ABySS: A parallel assembler for short read sequence data. Genome Res. 19:1117–1123.

Stadler T. 2010. Sampling-through-time in birth-death trees. J Theor Biol. 267:396–404.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30:1312–1313.

Standley P.C. 1937. Studies of American plants, VII. Field Mus. Nat. Hist., Bot. ser. 17:155–224.

Swofford D.L. 2003. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods) Version 4. Sinauer Associates, Sunderland, Massachusetts. Available: https://paup.phylosolutions.com/.

Thomas S.K., Liu X., Du Z., Dong Y., Cummings A., Pokorny L., Xiang Q.-Y., Leebens-Mack J. 2021. Comprehending the Cornales: Phylogenetic reconstruction of the order using the Angiosperms 353 probe set. Am. J. Bot. In press.

Voloch C.M., Schrago C.G. 2012. Impact of the partitioning scheme on divergence times inferred from mammalian genomic data sets. Evol. Bioinform. 8:EBO. S9627.

Wang J., Tian S., Sun X., Cheng X., Duan N., Tao J., Shen G. 2020. Construction of Pseudomolecules for the Chinese Chestnut (*Castanea mollissima*) Genome. G3-GENES GENOM GENET. 10:3565–3574.

Weitemier K., Straub S.C.K., Cronn R.C., Fishbein M., Schmickl R., McDonnell A., Liston A. 2014. Hyb-Seq: Combining Target Enrichment and Genome Skimming for Plant Phylogenomics. Appl. Plant Sci. 2:1400042.

Wen J., Shi S. 1999. A phylogenetic and biogeographic study of Hamamelis (Hamamelidaceae), an eastern Asian and eastern North American disjunct genus. Biochem. Syst. Ecol. 27:55–66.

Wiens J.J., Morrill M.C. 2011. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. Syst. Biol. 60:719–731.

Xiang Q.-Y., Crawford D.J., Wolfe A.D., Tang Y.-C., DePamphilis C.W. 1998. Origin and Biogeography of *Aesculus* L. (Hippocastanaceae): A Molecular Phylogenetic Perspective. Evolution. 52.

Xie L., Yi T.-S., Li R., Li D.-Z., Wen J. 2010. Evolution and biogeographic diversification of the witch-hazel genus (*Hamamelis* L., Hamamelidaceae) in the Northern Hemisphere. Mol. Phylogenet. Evol. 56:675–689.

Yu G., Lam T.T.-Y., Zhu H., Guan Y. 2018. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using *Ggtree*. Mol. Biol. Evol. 35:3041–3043.

Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC bioinform. 19:153.

Zhang Z., Lu A. 1995. Hamamelidaceae: geographic distribution, fossil history and origin. Chinese Science Abstracts Series B. 6:37.

Zhou W., Xiang Q.-Y., Wen J. 2020. Phylogenomics, biogeography, and evolution of morphology and ecological niche of the eastern Asian–eastern North American *Nyssa* (Nyssaceae). J. Syst. Evol. 58:571–603.

Table 1. Summary information of Angiosperm 353 gene data obtained for *Castanea* and *Hamamelis*, including the loci number, total MSA length, average length per locus, segregating sites number and percentage, parsimony informative sites number and percentage, and concatenation-based tree and species tree topologies.

Genus	Matrix	Gene groups	Total number of loci	Total MSA length (bp)	Average length per locus (bp)	Total hypervariable sites removed (per gene)	Number of segregating sites (%)	Number of parsimony informative sites (%)	Concatenation-based tree/ASTRAL-III/SVDQuartets
Castanea	untrimmed *consensus*	orthologs*	333	1277569	3836.5	--	151735 (11.9%)	40772 (3.2%)	**Top1/Top1/Top1**
	gappy trimmed *consensus*	orthologs	333	878334	2637.6	--	113444 (12.9%)	34209 (3.9%)	**Top1/Top1/Top1**
	PPD Trimmed consensus	orthologs*	333	823951	2474.3	54409 (163.4)	54410 (6.6%)	15007 (1.8%)	**Top1/Top1/Top1**
		paralogs	11	--	--	--	--	--	--
		All genes	344	842054	2447.8	55220 (160.5)	46206 (5.5%)	12295 (1.5%)	**Top1/Top1/Top1**
	degenerated	orthologs*	296	718887	2428.7	41356 (139.7)	40148 (5.6%)	10638 (1.5%)	**Top1/Top1/Top1**
		paralogs*	48	123167	2566.0	13864 (288.8)	6058 (4.9%)	1657 (1.3%)	Top1/Top2/Top1
Hamamelis	untrimmed *consensus*	orthologs*	344	760767	2211.5	--	68892 (8.9%)	14756 (2.3%)	Top5/Top4/Top4
	gappy trimed *consensus*	orthologs	344	607159	1765.0	--	54136 (8.9%)	14110 (2.3%)	Top5/Top4/Top4
	PPD Trimmed consensus	orthologs*	344	566772	1647.6	38050 (110.6)	26177 (4.6%)	7043 (1.2%)	**Top3/Top3/Top3**
		paralogs	2	--	--	--	--	--	--
		All genes	346	568476	1643.0	33743 (97.5)	22130 (3.9%)	6292 (1.1%)	**Top3/Top3/Top3**
	degenerated	orthologs*	319	514351	1612.4	29628 (92.8)	20185 (3.9%)	5693 (1.1%)	**Top3/Top3/Top3**
		paralogs*	27	54125	2004.6	4115 (153.9)	1945 (3.6%)	599 (1.1%)	Top4/Top5/Top5

Notes: The "untrimmed consensus matrices" were the orthologs directly generated by HybPiper, the "gappy trimmed consensus matrices" were the orthologs generated by HybPiper trimmed by s6 of part 2 of PPD. the "consensus matrices" were generated with HybPiper with all the PPD trimming steps, and the "degenerated" datasets were all generated with all steps of PPD. The "*consensus*" matrix contains sequences with heterozygous sites represented by the most frequent base; "*degenerated*" matrix contains sequences with heterozygous sites represented by the IUPAC ambiguity codes. Top 1: *Castanea,* ((EA), (ENA,

Europe)); Top 2: *Castanea,* ((EA, Europe), (ENA)); Top 3: *Hamamelis,* ((*H. mollis*, (*H. japonica*, ENA)); Top 4: *Hamamelis*, ((*H. japonica*, *H. mollis*), ENA); Top 5: *Hamamelis*, (*H. japonica*, (*H. mollis*, ENA)). MSA: Multiple Sequence Alignment. The matrices indicated by an asterisk were used for divergence time dating comparisons. Dash line indicates data not examined. Bolded font indicates the tree topologies are concordant in concatenation-based tree and species trees.

Table 2. The number of genes with different hits using BLAST against *Castanea* genome

Identified genes	PPD			HybPiper		
	multiple hits	single hit	no hit	multiple hits	single hit	no hit
Paralogs	31 (77.5%)	9 (22.5%)	0 (0.0%)	4 (36.4%)	6 (54.5%)	1 (9.1%)
Orthologs	46 (15.1%)	255 (83.9%)	3 (1.0%)	73 (21.9%)	258 (77.5%)	2 (0.6%)

FIGURE CAPTIONS

Figure 1. Concepts of ortholog-paralog and the key steps of workflow of HybPiper for target enrichment data analysis. a) Illustration of orthologs and paralogs. Orthologs are generated by speciation events, while paralogs are generated by gene duplication events. Speciation events are indicated by S1 and S2. b) The workflow of HybPiper, including reads mapping using BLASTx or BWA, de novo assembly of contigs using SPAdes, and paralogs detection.

Figure 2. Illustrated putative paralogs detection (PPD) pipeline analytical workflow. The flowchart shows the basic PPD functions, including two major parts: (1) generating "*degenerated*" sequences and (2) generating well-trimmed matrix and detecting paralogous genes consisting of eight steps (s1-s8). See details in pipeline description. MSA: multiple sequence alignment; Sp: species; OG: Outgroup.

Figure 3. Comparison of concatenation-based trees of *Castanea* and *Hamamelis* resulting from phylogenetic analyses of PPD (degenerated) and HypPiper (consensus) data of orthologs. a) degenerated, orthologs in *Castanea* from PPD. b) degenerated, orthologs in *Hamamelis* from PPD. c) PPD-trimmed consensus, orthologs in *Castanea* from HybPiper. d) PPD-trimmed consensus, orthologs in *Hamamelis* from HybPiper. e) untrimmed consensus, orthologs in *Castanea* from HybPiper. f) untrimmed consensus, orthologs in *Hamamelis* from HybPiper. All analyses were performed using the IQ-TREE partitioned by genes. The topologies of a) and b) are identical to the best merged partitioned concatenation-based trees from IQ-TREE and unpartitioned concatenation-based trees from RAxML and MrBayes (Supplementary Figs. S2 -

S7; available on Dryad). *Quercus castaneifolia* was used as the outgroup for *Castanea* while *Parrotiopsis jacquemontiana* and *Fothergilla gardenii* were used as outgroups of *Hamamelis*.

Figure 4. Coalescent-based species trees of *Castanea* and *Hamamelis* reconstructed using the orthologous gene data from PPD with ASTRAL and SVDQuartets. Numbers on the branches are support values from ASTRAL-III (the fractions of quartet trees supporting the node) and SVDQuartets (bootstrap support), respectively. *Quercus castaneifolia* was used as the outgroup of *Castanea* while *Parrotiopsis jacquemontiana* and *Fothergilla gardenii* were used as outgroups of *Hamamelis*.

Figure 5. Results of divergence times (median) of *Castanea* (left column) and *Hamamelis* (right column) estimated using different Hyb-Seq data. a) and b), the results are from PPD-generated orthologous data. c) and d), results are from "*consensus*" data through HybPiper and PPD trimming steps. e) and f), the results are from PPD-generated paralogous data. g) and h) results are from "*consensus*" data without any trimming steps. The number at each node represents the median divergence time, while the node bars represent its 95% HPD. Circle 1 represents the fossil calibration of *Castanea*, while circle 2 represents the fossil calibration of *Hamamelis*. All topologies were drawn by ggtree in R.

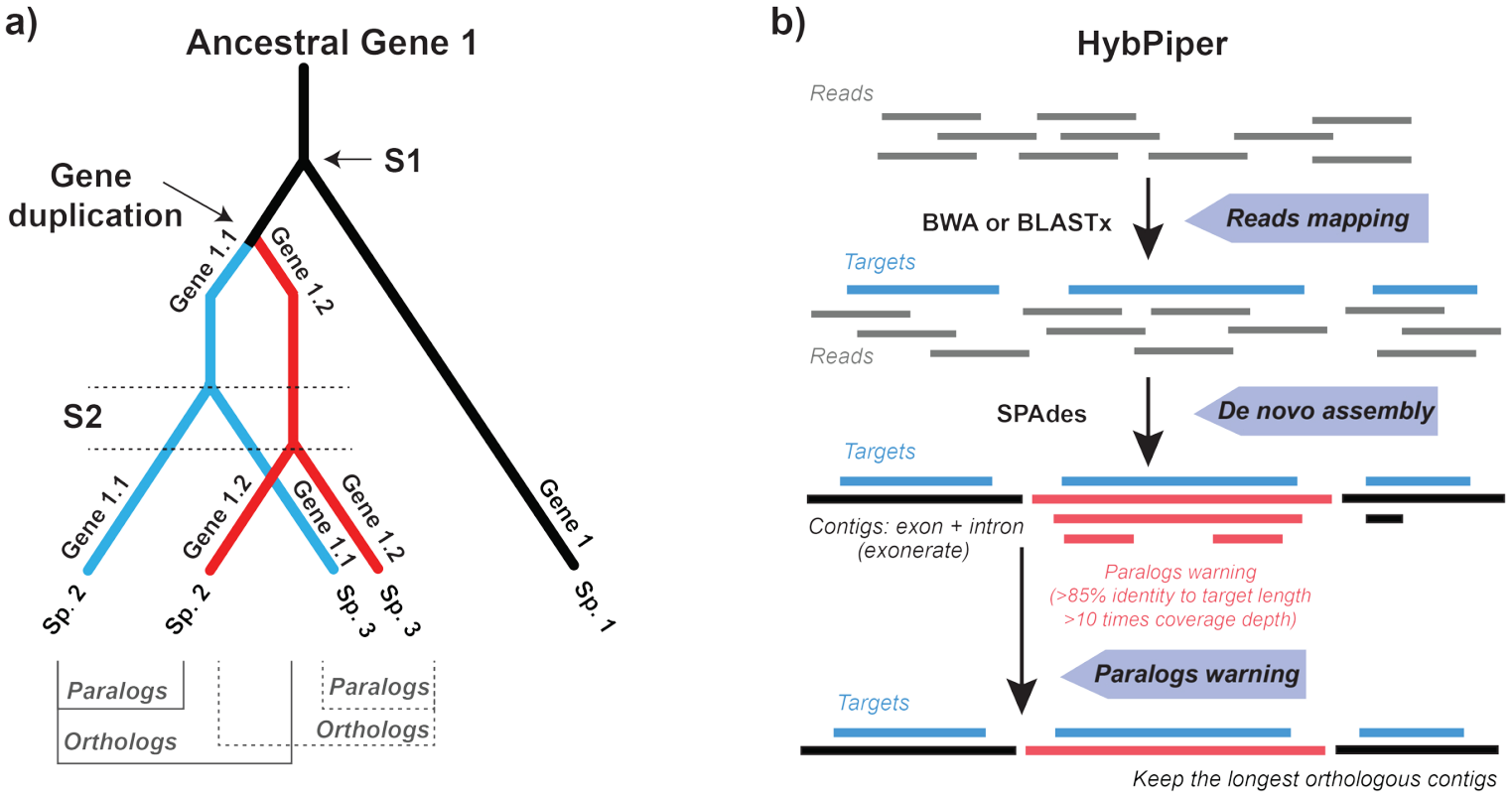

FIGURE 1

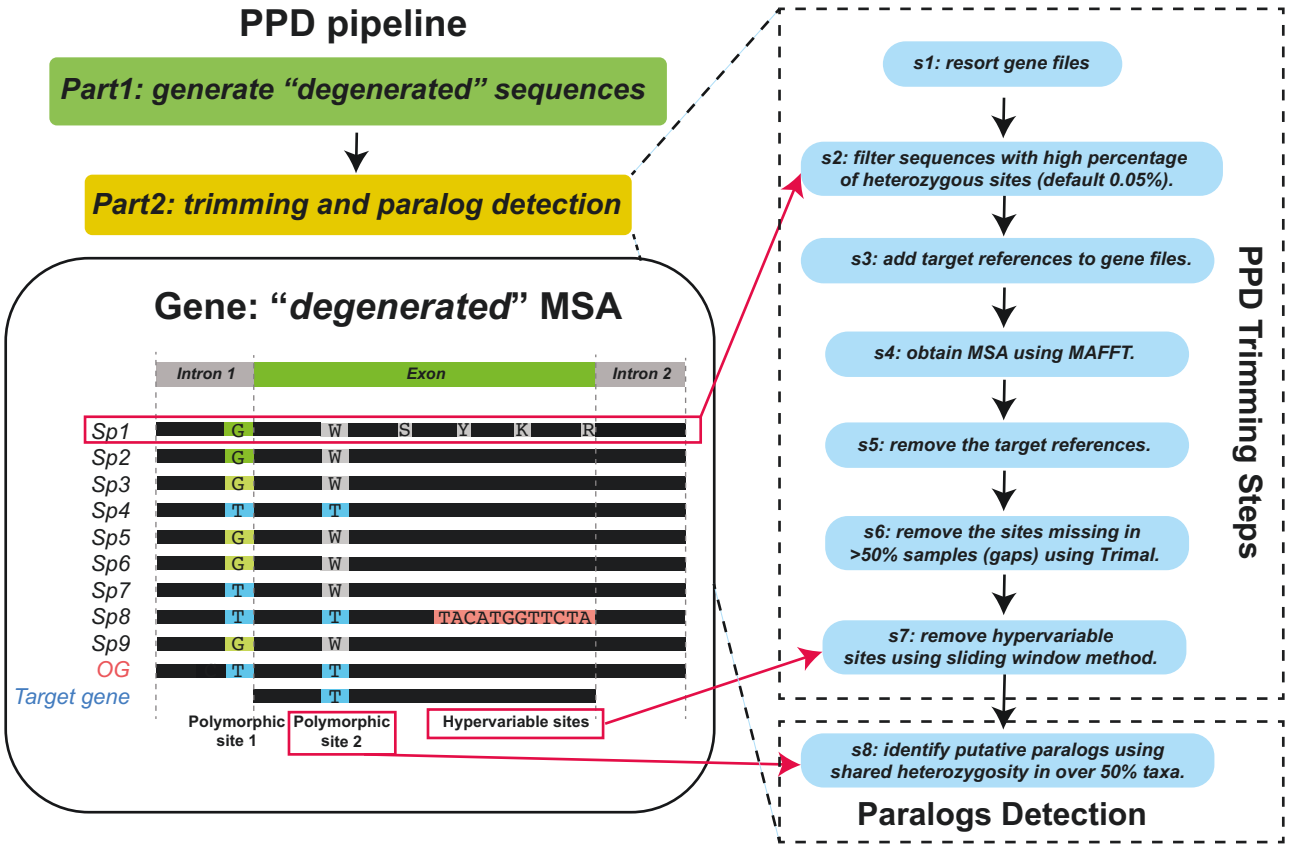

FIGURE 2

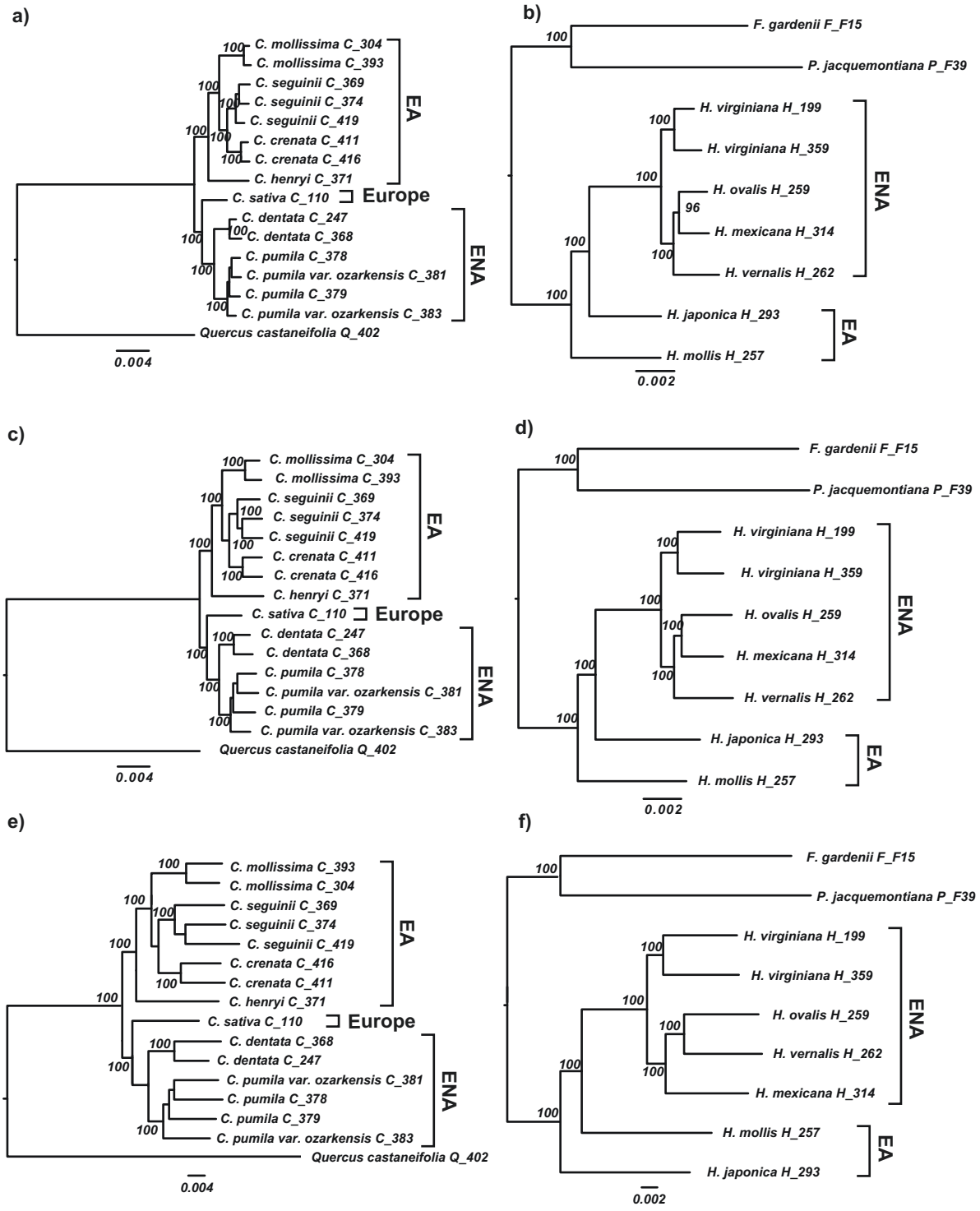

FIGURE 3

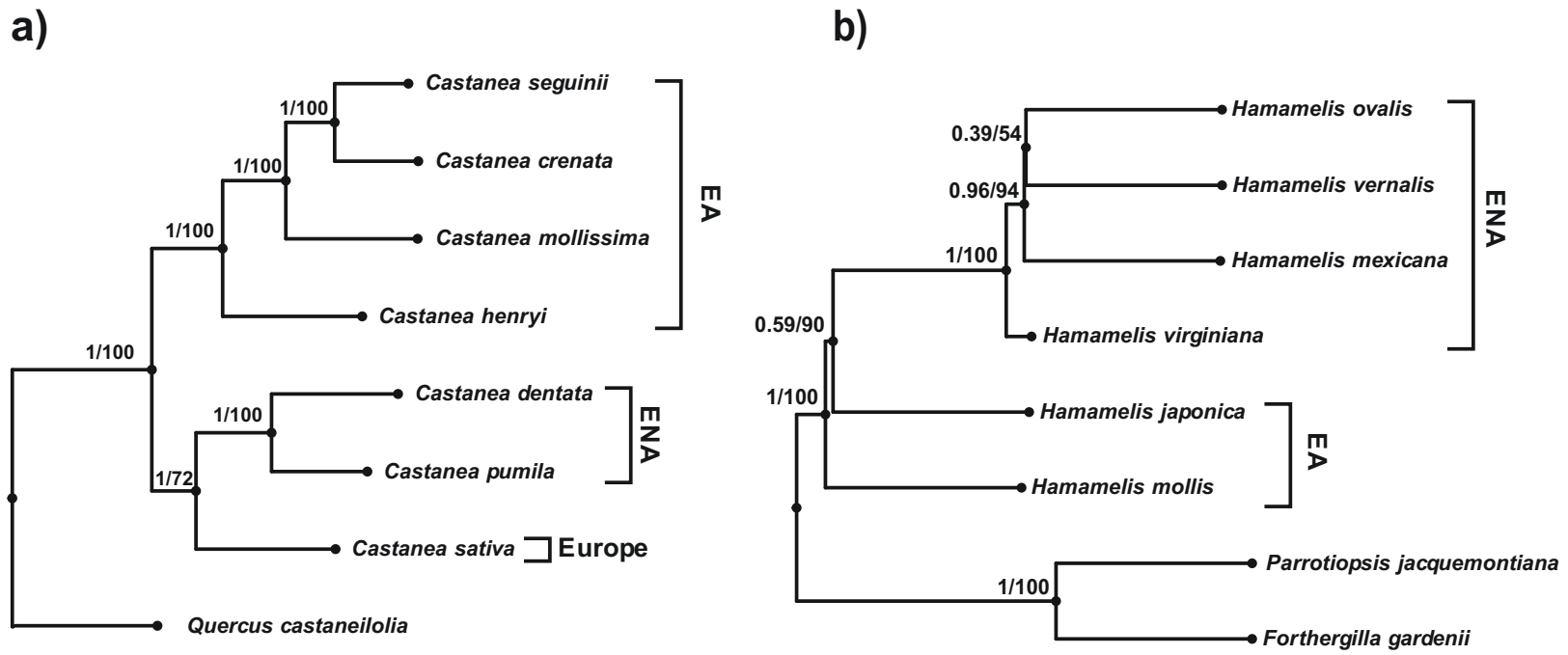

FIGURE 4

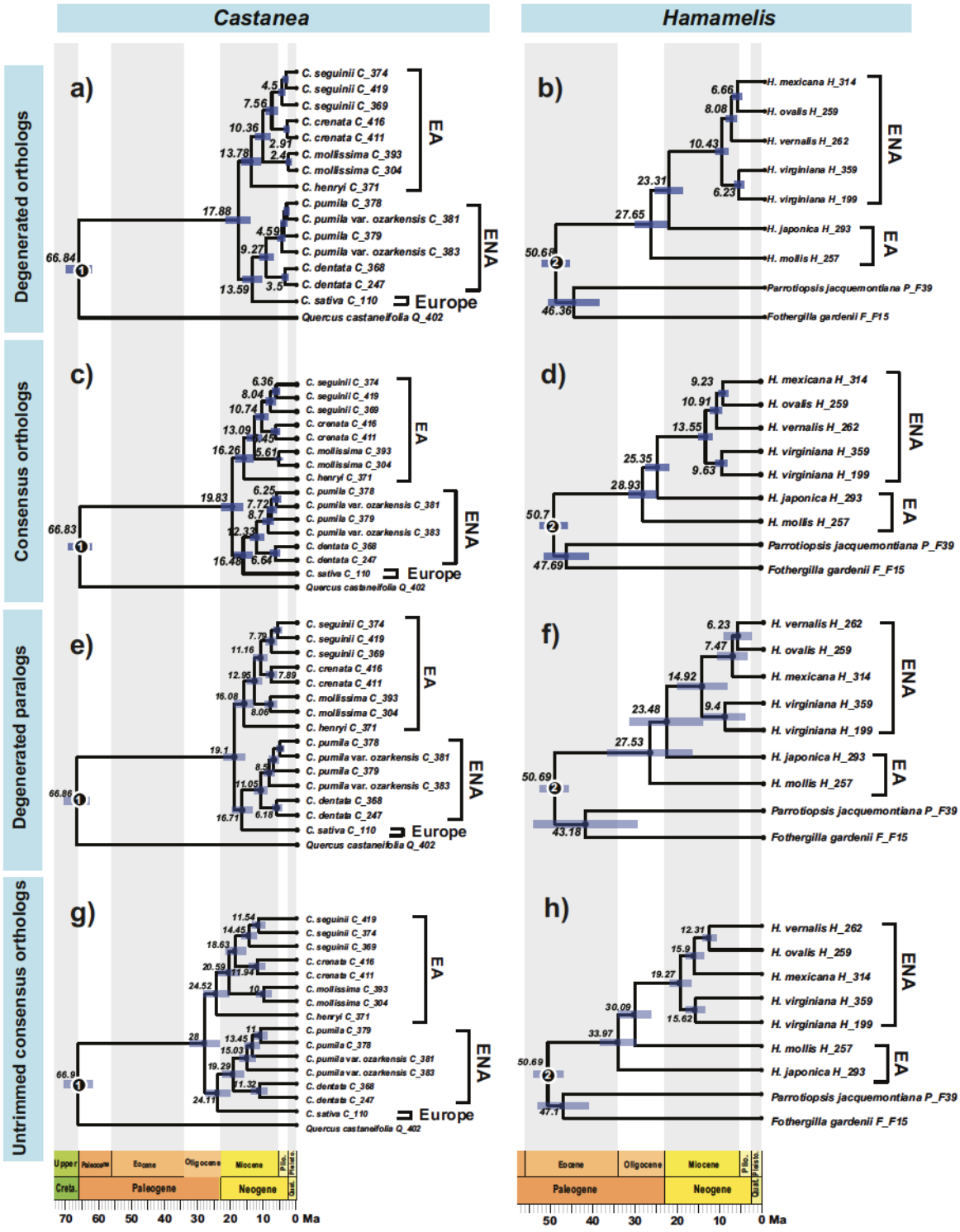

FIGURE 5