1 A New Paralog Removal Pipeline Resolves Conflict between RAD-seq and Enrichment

2 Wenbin Zhou^{1*}, John Soghigian², Qiu-yun (Jenny) Xiang^{1*}

3 *1 Department of Plant and Microbial Biology, North Carolina State University, Raleigh, NC*

4 *27965, USA*

5 *2 Department of Entomology and Plant Pathology, North Carolina State University, Raleigh, NC*

6 *27965, USA*

7

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

9 ABSTRACT

10 Target enrichment and RAD-seq are well-established high throughput sequencing
 11 technologies that have been increasingly used for phylogenomic studies, and the choice between
 12 methods is a practical issue for plant systematists studying the evolutionary histories of
 13 biodiversity of relatively recent origins. However, few studies have compared the congruence and
 14 conflict between results from the two methods within the same group of organisms, especially in
 15 plants, where extensive genome duplication events may complicate phylogenomic
 16 analyses. Unfortunately, currently widely used pipelines for target enrichment data analysis do not
 17 have a vigorous procedure for remove paralogs in Hyb-Seq data. In this study, we employed RAD-
 18 seq and Hyb-Seq of Angiosperm 353 genes in phylogenomic and biogeographic studies of
 19 *Hamamelis* (the witch-hazels) and *Castanea* (chestnuts), two classic examples exhibiting the well-
 20 known eastern Asian-eastern North American disjunct distribution. We compared these two
 21 methods side by side and developed a new pipeline (PPD) with a more vigorous removal of
 22 putative paralogs from Hyb-Seq data. The new pipeline considers both sequence similarity and
 23 heterozygous sites at each locus in identification of paralogous. We used our pipeline to construct
 24 robust datasets for comparison between methods and downstream analyses on the two genera. Our
 25 results demonstrated that the PPD identified many more putative paralogs than the popular method
 26 HybPiper. Comparisons of tree topologies and divergence times showed significant differences
 27 between data from HybPiper and data from our new PPD pipeline, likely due to the error signals
 28 from the paralogous genes undetected by HybPiper, but trimmed by PPD. We found that
 29 phylogenies and divergence times estimated from our RAD-seq and Hyb-Seq-PPD were largely
 30 congruent. We highlight the importance of removal paralogs in enrichment data, and discuss the
 31 merits of RAD-seq and Hyb-Seq. Finally, phylogenetic analyses of RAD-seq and Hyb-Seq

A NEW PARALOG REMOVAL PIPELINE

32 resulted in well-resolved species relationships, and revealed ancient introgression in both genera.
 33 Biogeographic analyses including fossil data revealed a complicated history of each genus
 34 involving multiple intercontinental dispersals and local extinctions in areas outside of the taxa's
 35 modern ranges in both the Paleogene and Neogene. Our study demonstrates the value of additional
 36 steps for filtering paralogous gene content from Angiosperm 353 data, such as our new PPD
 37 pipeline described in this study. [RAD-seq, Hyb-Seq, paralogs, *Castanea*, *Hamamelis*, eastern
 38 Asia-eastern North America disjunction, biogeography, ancient introgression]

INTRODUCTION

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. Landis et al. 2018; Lu et al. 2018; Dong et al. 2019; One Thousand Plant Transcriptomes Initiative 2019; and see reviews in Lemmon and Lemmon 2013; McCormack et al. 2013; Zimmer and Wen 2015; Chanderbali et al. 2017; Ma et al. 2017; Funk 2018; Lewin et al. 2018; Dodsworth et al. 2019). Among these HTS technologies, restriction site associated DNA sequencing (RAD-seq) and target enrichment (Hyb-Seq in plants or sequence capture and ultraconserved elements, UCEs, in animals) are highly promising and are increasingly used for phylogenomic study of lineages across different evolutionary phylogenetic timescales (e.g. Eaton and Ree 2013; Fernández-Mazuecos et al. 2018; Zhou et al. 2018; Du et al. 2019; MacGuigan and Near 2019 for RAD-seq data; Faircloth et al. 2013; McCormack et al. 2013; Leache et al. 2015; Lévillé-Bourret et al. 2018; Gaynor et al. 2020 for Target enrichment data).

RAD-seq and target enrichment methods each have their own strengths and weaknesses, which have been previously discussed (reviewed in Lemmon and Lemmon 2013; McCormack et al. 2013; Zimmer and Wen 2015; Harvey et al. 2016; Dodsworth et al. 2019). The main differences involve the portion of the genome captured, library preparations procedures, pipelines for data analyses, and cost. RAD-seq approaches are relatively easy and have a relatively lower cost for the laboratory work and provide deep coverage of numerous loci associated with the cutting sites of the restriction enzyme(s) employed. Furthermore, no prior knowledge of gene sequences is

A NEW PARALOG REMOVAL PIPELINE

62 required (Peterson et al. 2012; Leache et al. 2015; Andrews et al. 2016; Harvey et al. 2016;
63 Manthey et al. 2016; Dodsworth et al. 2019). In contrast, target enrichment methods cost more per
64 sample and have relatively more intensive lab work, and the number of loci is generally limited to
65 a targeted set of highly conserved genomic regions (and their flanking areas), often protein coding
66 genes. Prior knowledge of target sequences is needed to design the probes of targeted genes, either
67 from the organism of interest, or a related species. However, the target enrichment methods
68 generally have greater repeatability between experiments and between labs if the same probes are
69 used (Harvey et al. 2016), and typically generate relatively longer sequences at each locus.
70 Therefore, data from target enrichment are a lasting and amplifiable resource for comparative
71 studies at multiple taxonomic scales. Additionally, as target enrichment is often based on a known
72 set of conserved regions or genes, researchers may also know better the function and genomic
73 distribution of the genes targeted. Conversely, RAD-seq data have relatively low repeatability
74 between experiments, labs, and taxa as a result of variation in DNA samples and experimental
75 procedures and reagents (DaCosta and Sorenson 2014; Flanagan and Jones 2018; Malinsky et al.
76 2018). In addition, target enrichment data are suitable to phylogenomic studies of both deep and
77 shallow phylogenetic divergence because the data contain both conserved coding sequences and
78 their flanking variable sequences, whereas RAD-seq data is traditionally thought to be more
79 appropriate for studies of shallow phylogenetic divergence (e.g., studies within a genus or species,
80 and sometimes in relatively young families) due to allele dropout effects (Arnold et al. 2013;
81 Gautier et al. 2013; Eaton 2014, and reviewed in Leaché and Oaks 2017). However, recent studies
82 have shown that the allele-drop out (missing data) in RAD-seq data is usually systematic (or
83 hierarchical) and phylogenetically informative (DaCosta and Sorenson 2016). Recent studies using
84 RAD-seq have resolved phylogenies of lineages diverging during the Eocene and even the late

Cretaceous (Eaton et al. 2016; Hou et al. 2016; Leaché and Oaks 2017; Lecaudey et al. 2018; Zhou et al. 2018; MacGuigan and Near 2019; Du et al. 2020; Mu et al. 2020).

For a phylogenetic study of a lineage at the phylogenetic timescales applicable to both target enrichment and RAD-seq approaches, e.g., building species level phylogenies, researchers may choose one of these approaches with the belief that either option would produce robust data for correct inference of phylogeny and divergence times. However, we have little knowledge about the level of congruence between the two data types and the impact of data analysis procedures. Few studies have made efforts to compare results derived from target enrichment and RAD-seq. At present, only a few studies on animals have applied both methods to allow a direct comparison of the phylogenetic utilities between the target enrichment data and RAD-seq data and to evaluate their congruences (Leache et al. 2015; Harvey et al. 2016; Manthey et al. 2016). Harvey et al. (2016) and Manthey et al. (2016) found that the topologies from RAD-seq and sequence capture were similar for lineages of birds that originated in the late Miocene or more recently, but with longer branch lengths for sequence capture. However, Leache et al. (2015), in an analysis of lizard lineage dating back to the Eocene, found differences in concatenated gene topologies and the species relationships between the target enrichment data and RAD-seq data. These findings are limited to only a few case studies, in which how the two datasets behave in subsequent phylogeny-based and branch length-dependent comparative analyses, such as divergence time dating and range evolution, was not assessed. To our knowledge, this is the first study comparing the phylogenetic outcomes between data from Hyb-Seq, the target enrichment approach developed for plant phylogenomics by Weitemier et al. (2014), and data from RAD-seq.

A NEW PARALOG REMOVAL PIPELINE

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 genes (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 genes targeted 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 targeted 353 genes among the diverse angiosperm genomes poses a concern on possible prevalence of paralogs in the Hyb-Seq data. It is unknown if the current pipelines developed for analyses of target enrichment data have the ability to reliably exclude paralogous gene copies in the Hyb-Seq 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. 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). Therefore, excluding paralogous comparisons in phylogenetic studies of species relationships 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

Fig. 1. Concepts of ortholog-paralog and the key steps of workflow of current bioinformatics programs 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 (Johnson et al. 2016), including reads mapping using BLASTx or BWA (Li and Durbin 2010), de novo assembly of contigs using SPAdes (Bankevich et al. 2012) and paralogs (in red) detection. c) The workflow of PHYLUCE (Faircloth 2016) and SECAPR (Andermann et al. 2018a), including de novo assembly of contigs from reads using SPAdes (in PHYLUCE) and Abyss (Simpson et al. 2009) (in SECAPR), respectively, contigs mapping using LASTZ (Harris 2007), and detection

A NEW PARALOG REMOVAL PIPELINE

of paralogs (in red), followed by removal of detected paralogs. All three methods detect putative paralogs by observation of more than one unique contigs in an individual mapped to the same reference gene with >85% (HybPiper) or >80% (PHYLUCE and SECAPR) sequence identity or the same contig mapped to more than one reference genes.

A sequence similarity-based approach for calling paralogous genes, such as those implemented in the three popular pipelines mentioned above, may be sufficient for phylogenetic studies using custom designed probes based on orthologous sequences encompassing the 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 a supplementary criterion to the sequence similarity for detecting and removing the problematic paralogous gene data from Hyb-Seq by examining heterozygous sites within and among individuals in the aligned sequences (Fig. 2). Low heterozygosity 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). Even at the shallow level of phylogenetic divergence (e.g. population genetics), high heterozygosity within a locus may also be attributed to paralogs (Hohenlohe et al. 2011; Harvey et al. 2016; McKinney et al. 2017). Additionally, a high number of heterozygous sites of a locus within an individual can be used as an indicator for gene duplication events or previously undetected polyploidy (Medina et al. 2019). Most enrichment pipelines do not by default consider the sequence variants of contigs/loci (details

in Supplementary Information, available on Dryad). In contrast, the methods developed for analyses of RAD-seq data are usually stricter and use more parameters to filter putative paralogs (Harvey et al. 2016). For example, PyRAD (Eaton 2014) not only uses the parameter of clustering based on sequence similarity within each locus, but also uses two other parameters to filter sequences with high heterozygous sites and with high percentage of shared heterozygous sites (Eaton 2014; McKinney et al. 2017). Therefore, we modified HybPiper (Fig. 2) to include analyses of the heterozygosity parameters for further filtering of putative paralogs from enrichment data.

Fig. 2. The putative paralogs detection (PPD) pipeline analytical workflow. The flowchart shows the basic PPD functions, including two major steps: first, generating a "degenerated" sequences; second, generating well-trimmed matrix and detecting paralogous genes. The "well-trimmed" process (s1 to s7) include filtering out sequences with high percentage of heterozygous sites (>5%), constructing Multiple Sequence Alignment (MSA), trimming MSA by removing sites missing in >50% samples (gaps) and hypervariable sites by a sliding window trimming

A NEW PARALOG REMOVAL PIPELINE

method. The paralog detection step (s8) can detect additional paralogs indicated by presence of heterozygous site(s) present in >50% samples in the MSA. The “*degenerated*” sequence means the sequence were coded by IUPAC ambiguity codes. See details in Supplementary Information, available on Dryad. MSA: multiple sequence alignment; Sp: species; OG: Outgroup.

In order to compare Hyb-Seq and RAD-seq, as well as to evaluate how putative paralog detection impacted the results of a Hyb-Seq analysis, we conducted extensive analyses of Hyb-Seq data from the Angiosperm 353 kit and RAD-seq data from two genera: *Castanea* (Chestnuts of Fagaceae) and *Hamamelis* (Witch-hazel of Hamamelidaceae). From both genera, we inferred species relationships, divergence times, and biogeographic histories and compared the results to gain insights into the congruence between Angiosperm 353 and RAD-seq data. These two genera are common elements of the deciduous forests of the northern hemisphere and discontinuously distributed in eastern Asia (EA) and eastern North America (ENA); the study of which allow us to gain insights into the origin and evolution of this well-known floristic disjunction, one of the first order biodiversity patterns in the Northern Hemisphere that has been extensively studied (Wolfe 1980; Boufford and Spongberg 1983; Tiffney 1985a; Xiang et al. 1998b, 2000; Wen 1999; Donoghue and Smith 2004; Wen et al. 2010).

The chestnut genus *Castanea* Miller (Fagaceae) includes seven tree species, restricted to EA and ENA, with one species endemic to 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* 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 and sect. *Balanocastanon* and sect. *Hypocastanon* did not form a sister group. Authors of these studies also hypothesized an eastern Asia origin of *Castanea* in the Eocene, followed by range expansion to Europe and North America in the mid Eocene, as an example of the “Out of Asia” pattern (Donoghue and Smith 2004).

The witchhazel genus *Hamamelis* L. (Hamamelidaceae) is also a small genus consisting of four or 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* have been 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.

A NEW PARALOG REMOVAL PIPELINE

In this study, the first goal is to evaluate the relative merits of RAD-seq and target enrichment methods for addressing within-genus-level phylogenetic researches, through gene tree and species tree comparisons, as well as divergence time comparisons. Our second goal is to generate a new pipeline for detecting more putative paralogs for downstream enrichment data analyses rather than just using the paralog detection method from HybPiper. We compared multiple sequence alignment (MSA) matrices generated using our pipeline with the MSA matrices generated from HybPiper using different character coding and trimming methods (See Methods). We further compared the phylogenies and divergence times resulted from analyses of the different data matrices to determine the effects of heterozygotes and paralogs on phylogenetic and divergence time analyses. Finally, we addressed longstanding questions regarding the evolutionary history of *Castanea* and *Hamamelis* and the EA-ENA floristic disjunction. Combining the molecular and fossil records, we conducted rigorous phylogenomic and comprehensive biogeographic analyses to uncover their geographic origins, migration routes, and past hybridization and to gain new insights into the formation of the EA-ENA disjunct phytogeographic pattern.

MATERIALS AND METHODS

Data Generation

DNA samples preparation.—A total of 58 individuals were sampled and sequenced. The samples consisted of 30 plants of *Castanea* and 28 plants of *Hamamelis*, including two and three outgroup individuals, respectively (Table 1). 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*. The sampling included all species of the two ingroup genera recognized in modern floras. Leaf samples were collected from the fields or plants grown in arboreta or botanical gardens (Table 1). 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. All the vouchers were deposited in the herbarium of North Carolina State University (acronym NCSC).

Total genomic DNAs were extracted from leaf samples using the CTAB protocol (Doyle 1991) with modification described in Cullings (1992) and Xiang et al. (1998a). 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 quantity and quality 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).

A NEW PARALOG REMOVAL PIPELINE

Table 1 Sample information.

Taxa and individual #	RAD-seq #	Hyb-Seq #	State-County	Voucher	Herbarium
<i>Castanea crenata</i> Siebold & Zucc.	CC_31	411	Modong-myeon, Korea	Joongku Lee # CNU DG039	NA
<i>Castanea crenata</i>	CC_32	416	Modong-myeon, Korea	Joongku Lee # CNU BH 025	NA
<i>Castanea crenata</i>	CC_33	NA	JC Arboretum, NC, USA	Wenbin Zhou # Xianglab204	NCSC
<i>Castanea henryi</i> (Skan) Rehder & E.H. Wilson	CH_40	NA	Kunming, China	Shaolin Tan # KBG2017005	NCSC
<i>Castanea henryi</i>	CH_41	371	Kunming, China	Shaolin Tan # KBG2017006	NCSC
<i>Castanea mollissima</i> Blume	CM_42	NA	Arnold Arboretum, MA, USA	# 269-2011*A	A
<i>Castanea mollissima</i>	CM_43	304	UW Botanic Gardens, WA, USA	# 2338-37*A	A
<i>Castanea mollissima</i>	CM_44	NA	Kunming, China	Shaolin Tan # KBG2017007	NCSC
<i>Castanea mollissima</i>	CM_45	393	Hangzhou, China	Yunpeng Zhao et al. # HZU00115	HZU
<i>Castanea mollissima</i>	CM_55	NA	Arnold Arboretum, MA, USA	# 12924*A	A
<i>Castanea mollissima</i>	CM_56	NA	Arnold Arboretum, MA, USA	# 12924*B	A
<i>Castanea seguinii</i> Dode	CSE_46	374	Hangzhou, China	Yunpeng Zhao et al. # HZU00123	HZU
<i>Castanea seguinii</i>	CSE_47	NA	Hangzhou, China	Yunpeng Zhao et al. # HZU00423	HZU
<i>Castanea seguinii</i>	CSE_38	419	Hangzhou, China	Yunpeng Zhao et al. # HZU00422	HZU
<i>Castanea seguinii</i>	CSE_39	369	Kunming, China	Shaolin Tan # KBG2017004	NCSC
<i>Castanea dentata</i> (Marshall) Brokh.	CD_34	247	Arnold Arboretum, MA, USA	# 224-97*A	A
<i>Castanea dentata</i>	CD_36	NA	Giles, VA, USA	Mark Whitten # DOB 2512	FLAS
<i>Castanea dentata</i>	CD_37	368	Worcester, MA, USA	Mark Whitten # DOB 3225	FLAS
<i>Castanea pumila</i> (L.) Mill.	CP_35	NA	Bibb, AL, USA	Mark Whitten # DOB 2117	FLAS
<i>Castanea pumila</i>	CP_48	NA	Putnam, FL, USA	Mark Whitten # DOB 4	FLAS
<i>Castanea pumila</i>	CP_49	378	Putnam, FL, USA	Lucas Majure # 4755	FLAS
<i>Castanea pumila</i>	CP_50	379	Macon, NC, USA	Mark Whitten # DOB 3111	FLAS
<i>Castanea pumila</i>	CP_51	NA	Arnold Arboretum, MA, USA	# 21486*A	A
<i>Castanea pumila</i> var. <i>ozarkensis</i> (Ashe) A.E. Murray	CPO_52	NA	Arnold Arboretum, MA, USA	# 336-55*B	A
<i>Castanea pumila</i> var. <i>ozarkensis</i>	CPO_53	381	Arnold Arboretum, MA, USA	# 227-97*A	A
<i>Castanea pumila</i> var. <i>ozarkensis</i>	CPO_54	383	Arnold Arboretum, MA, USA	# 366-2010*B	A
<i>Castanea sativa</i> Mill.	CS_57	NA	Jammu and Kashmir	Alfred Linn Bogle et al. # 151	NA

ZHOU ET AL.

<i>Castanea sativa</i>	CS_58	NA	Northwest Himalaya	Ralph Randles Stewart # 16980	NA
<i>Castanea sativa</i>	NA	110	Ladenburg, Germany	Yalu Ru # 18060101	NA
<i>Fagus grandifolia</i> Ehrh.	FAG_59	NA	JC Arboretum, NC, USA	Wenbin Zhou # Xianglab195	NCSC
<i>Quercus castaneifolia</i> C.A. Mey.	QC_60	402	Arnold Arboretum, MA, USA	# 239-38*D	A
<i>Hamamelis japonica</i> Siebold & Zucc.	HJ_61	293	Arnold Arboretum, MA, USA	Wenbin Zhou # Xianglab269	NCSC
<i>Hamamelis japonica</i>	HJ_62	NA	UW Botanic Gardens, WA, USA	# 2275-37*A	NCSC
<i>Hamamelis japonica</i>	HJ_63	NA	UW Botanic Gardens, WA, USA	# X-407*A	NCSC
<i>Hamamelis mollis</i> Oliv.	HMO_66	257	Arnold Arboretum, MA, USA	# 698-94*A	A
<i>Hamamelis mollis</i>	HMO_67	NA	Kunming, China	Shaolin Tan # KBG2017011	NCSC
<i>Hamamelis mollis</i>	HMO_68	NA	Hangzhou, China	Chuan Chen # HZB006	HHBG
<i>Hamamelis mollis</i>	HMO_69	NA	Jiangxi, China	Ceming Tan # 831	KUN
<i>Hamamelis mollis</i>	HMO_70	NA	Hubei, China	Xiangyun Yang # 12CS5545	KUN
<i>Hamamelis ovalis</i> S.W. Leonard	HO_71	NA	JC Arboretum, NC, USA	Wenbin Zhou # Xianglab212	NCSC
<i>Hamamelis ovalis</i>	HO_72	259	Arnold Arboretum, MA, USA	# 114-2009*A	A
<i>Hamamelis ovalis</i>	HO_73	NA	Arnold Arboretum, MA, USA	# 114-2009*B	A
<i>Hamamelis ovalis</i>	HO_74	NA	USA	# 115-2009*A	A
<i>Hamamelis vernalis</i> Sarg.	HVE_75	NA	JC Arboretum, NC, USA	Wenbin Zhou # Xianglab199	NCSC
<i>Hamamelis vernalis</i>	HVE_77	262	Arnold Arboretum, MA, USA	# 336-89*A	A
<i>Hamamelis mexicana</i> Standl.	HME_64	314	UW Botanic Gardens, WA, USA	# 383-92*A	NCSC
<i>Hamamelis mexicana</i>	HME_78	NA	Ejido San Jose, Mexico	D. Seigler and F. A. Jones # DS-14198	LL
<i>Hamamelis virginiana</i>	HVI_79	NA	Arnold Arboretum, MA, USA	# 1337-85*A	A
<i>Hamamelis virginiana</i>	HVI_80	NA	Arnold Arboretum, MA, USA	# 1195-84*A	A
<i>Hamamelis virginiana</i>	HVI_81	NA	Carroll, NH, USA	Mark Whitten # DOB 3336	FLAS
<i>Hamamelis virginiana</i>	HVI_82	359	Macon, NC, USA	Mark Whitten # DOB 3055	FLAS
<i>Hamamelis virginiana</i>	HVI_83	NA	Worcester, MA, USA	Mark Whitten # DOB 3244	FLAS
<i>Hamamelis virginiana</i>	HVI_84	NA	Giles, VA, USA	Mark Whitten # DOB 2520	FLAS
<i>Hamamelis virginiana</i>	HVI_65	199	JC Arboretum, NC, USA	Wenbin Zhou # Xianglab220	NCSC

A NEW PARALOG REMOVAL PIPELINE

<i>Hamamelis x intermedia</i>	HI_85	NA	JC Arboretum, NC, USA	Wenbin Zhou # Xianglab203	NCSC
<i>Hamamelis vernalis</i>	HVE_86	NA	Arnold Arboretum, MA, USA	# 1201-66*A	A
<i>Fothergilla major</i> (Sims) Sweet	FOM_76	NA	JC Arboretum, NC, USA	Wenbin Zhou # Xianglab207	NCSC
<i>Fothergilla gardenii</i> L.	FOG_87	F15	Richmond, NC, USA	Zhechen Qi et al. (2015) # F15	NCSC
<i>Parrotiopsis jacquemontiana</i> (Decne.) Rehder	PJ_88	F39	Undocumented	Zhechen Qi et al. (2015) # F39	NCSC

Notes: NA means data are not available.

Library preparation for RAD-seq and sequencing.—Double digest RAD-seq (ddRAD) (Peterson et al. 2012) libraries were prepared following Pais et al. (2018, 2020) and Du et al. (2020) for all samples. Two restriction enzymes, *PstI*-HF and *MspI* were used to digest 350 ng of the total DNA from each sample (Table 1). We used the *PstI* adaptors and forward primer described in Pais et al. (2017) and Zhou et al. (2018), but modified the protocol with the *MspI* adaptor and reverse primer from Du et al. (2020). Libraries were pooled and size selected to 400-600 bp using a Pippin prep (Sage Science, Beverly, MA, USA) and verified by analysis on Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA). Size-selected libraries were enriched by PCR for 12 cycles, followed by purification with AMPure Beads before delivered to the Genomic Science Lab (GSL) at North Carolina State University for sequencing. Single-end sequencing (150 bp) from the end of the *PstI* cutting site was performed on an Illumina HiSeq 2500 platform.

Library preparation of Angiosperm 353 gene enrichment and sequencing.—Due to the relative high cost of sequence capture, we reduced the samples in this experiment to include 16 *Castanea*, eight *Hamamelis*, and four outgroups (Table 1), which also covers all species of the two genera. 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).

A NEW PARALOG REMOVAL PIPELINE

Locus Data Assembly and MSA Generation

RAD-seq data.—Raw sequence reads were analyzed in ipyrad v.0.7.28 (Eaton and Overcast 2020), summarized briefly here. The reads were first de-multiplexed and then trimmed to remove specific barcodes and adapters. Then, low quality reads (with Q score < 20) were trimmed and filtered. To remove the influence of DNA sequences from the plastid genome and microorganisms on phylogenetic inference, chloroplast genomes of *Castanea mollissima* (NC_014674, Jansen et al. 2011) and *Hamamelis mollis* (NC_037881, Dong et al. 2018) and the full bacterial and fungal database (<https://ftp.ncbi.nlm.nih.gov/refseq/release/>) were downloaded from NCBI as the references. We chose the “denovo-reference” function in ipyrad, which uses VSEARCH (Rognes et al. 2016) to cluster reads of the same locus and uses BWA (Li and Durbin 2009) and BEDTools (Quinlan and Hall 2010) to remove the reads/loci mapping to the plastid and microorganisms references. We selected a default threshold of 85% of similarity in sequences for reads clustering in ipyrad based on results in Du et al. (2020) for orthologs identification among plant species. We also used 2 as the maximum alleles per sample in the consensus sequence construction for each locus in a sample because both *Castanea* and *Hamamelis* are diploid organisms and a maximum of 50% samples sharing heterozygosity at a site (i.e., 0.5 shared heterozygosities in ipyrad) of a locus as the cut-off of orthology. The resulting read clusters were aligned across samples using MUSCLE (Edgar 2004). The loci were finally trimmed (5 bases on 3’ end) and filtered if they contained over 30 SNPs or 8 indels. To compare across different levels of missing data, we output matrices containing full-length loci in different minimum thresholds of samples using the parallel branching assembly function. We selected thresholds from at least 20% of samples present at a locus, to 80%, increasing in increments of 10%. We named these matrices based on this threshold,

e.g. M20, which contained all loci present in at least 20% of samples. All matrices have been uploaded to Dryad.

Hyb-Seq data.—All samples for Hyb-Seq 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 (Li and Durbin 2009), applying the SPAdes (Bankevich et al. 2012) to assemble reads into long contigs, and implementing the intronrate.py module to recover “intron” and “supercontig” (CDS + intron fragments) sequences. Finally, we generated multiple Hyb-Seq data matrices to assess the phylogenetic effects of character coding, data trimming, and paralogous genes in order to find the best dataset for comparison with the RAD-seq data and for assessing their impacts on divergence time. Specifically, we identified loci containing putative paralogs using our new pipeline, named putative paralog detection (PPD) (available on Github: https://github.com/Bean061/putative_paralog) and built the following matrices for comparisons: exon vs. supercontig matrices of orthologous loci, paralogous loci, or all genes (including the paralogous genes), respectively. The exon matrices contained the sequences of the coding regions only while the supercontig matrices contained sequences of both coding and their flanking regions of the three respective gene groups. The pipeline includes two major steps that occur after generating alignment matrices in HybPiper or other enrichment pipelines: first, “*degenerated*” sequences are generated from mapping reads back to contigs, which

A NEW PARALOG REMOVAL PIPELINE

334 applies the IUPAC ambiguity codes to capture the heterozygous sites in generating the exon and
 335 supercontig matrices of the three sets of aforementioned genes; and second, the matrix is trimmed
 336 of highly heterozygous sites, misaligned regions, and particularly gappy columns, as one trimming
 337 step which is shown in Figure 2 (see details in Supplementary Information available on Dryad).
 338 Thresholds for trimming are configurable by the user. To directly compare the impact of character
 339 coding, we compared the “*degenerated*” orthologous matrix generated through both steps of PPD
 340 with the orthologous data matrix built from HybPiper without degenerate character coding in the
 341 matrix (referred to as “*consensus*” matrices) and also trimmed these matrices with the second step
 342 of PPD (referred to as the “well-trimmed *consensus*” matrices). We also compared data from the
 343 Angiosperm 353 genes with the plastid genes captured by off-target in the Hyb-Seq experiment,
 344 and the combined genes from RAD-seq and Hyb-Seq (excluding plastid genes). In addition, we
 345 compared data from exon and supercontig, paralogs and orthologous, and applied three different
 346 trimming approaches (no trimming, automated trimming in Trimal, and our “well-trimmed”
 347 method in PPD) in producing these matrices (see details in Supplementary Information, available
 348 on Dryad). For all the Hyb-Seq data matrices below, if no trimming description is directly indicated,
 349 the data matrices were “well-trimmed” through PPD (PPD Part 2: s1-s7 in Fig. 2). All Hyb-Seq
 350 comparisons are listed in Table 2 (see details in Supplementary Information, available on Dryad).

351

352 **Table 2 Summary information of RAD-seq data, Angiosperm 353 data, chloroplast plastid data, and combined RAD-Hyb-Seq**
353 **data, including the loci number, total MSA length, average length per locus, segregating sites number and percentage,**
354 **parsimony segregating sites number and percentage, and concatenated gene tree and species tree topologies.**
355

Data source	Matrix (with BWA -k parameter)	gene regions	trimming & mapping methods	Homologs	Total loci	Total MSA length	average length per locus	Segregating sites number (%)	Parsimony informative sites (%)	species tree Astral	Species tree SVDQuartets	concatenated gene tree
<i>Castanea</i> RAD-seq	RAD-seq	NA	NA	NA	1290	135781	105.3	5261 (0.039)	2475 (0.018)	1	1	1
<i>Castanea</i> Hyb-Seq	consensus	exon	untrimmed	orthologs	330	204474	619.6181818	11555 (0.057)	4490 (0.022)	1	1	1
			well-trimmed	orthologs	330	194122	588.2484848	7069 (0.036)	2332 (0.012)	1	1	1
		supercontig	untrimmed	orthologs	333	878334	2637.6	113444 (0.129)	34209 (0.039)	1	1	1
			well-trimmed	orthologs	333	823951	2474.3	54410 (0.066)	15007 (0.017)	1	1	1
	degenerated (BWA19)	exon	auto-trimmed	All genes	339	185862	548.3	7866 (0.042)	2837 (0.015)	1	1	1
				orthologs	261	133482	511.4	5008 (0.038)	1693 (0.013)	1	1	1
				paralogs	89	57353	644.4	3061 (0.053)	1214 (0.021)	2	2	2
			well-trimmed	All genes	348	206695	594.0	5583 (0.027)	1625 (0.008)	1	1	1
				orthologs	248	138980	560.4	3894 (0.028)	1156 (0.008)	1	1	1
				paralogs	100	67715	677.2	1689 (0.025)	469 (0.007)	2	2	2
		supercontig	untrimmed	All genes	331	871309	2632.4	93393 (0.107)	26285 (0.03)	1	1	1
				orthologs	227	542954	2391.9	53162 (0.098)	13996 (0.026)	1	1	1
				paralogs	122	351468	2880.9	42986 (0.122)	13137 (0.037)	1	1	1
			auto-trimmed	All genes	333	360422	1082.3	21953 (0.061)	5318 (0.015)	1	2	1
				orthologs	258	263855	1022.7	15363 (0.058)	3913 (0.015)	1	1	1
				paralogs	92	106832	1161.2	7130 (0.067)	1662 (0.016)	1	2	2
			well-trimmed	All genes	349	846883	2426.6	43123 (0.051)	10870 (0.013)	1	1	1
				orthologs	233	548061	2352.2	29917 (0.055)	7832 (0.014)	1	1	1
				paralogs	116	298822	2576.1	13206 (0.044)	3038 (0.01)	1	1	1
	degenerated (BWA100)	supercontig	well-trimmed	All genes	344	842054	2447.8	46206 (0.055)	12295 (0.015)	1	1	1
				orthologs	296	718887	2428.7	40148 (0.056)	10638 (0.015)	1	1	1
				paralogs	48	123167	2566.0	6058 (0.049)	1657 (0.013)	1	2	1
<i>Castanea</i> Hyb-Seq Chloroplast	Chloroplast (BWA100)	NA	well-trimmed	NA	1	12647	12647.0	102 (0.008)	27 (0.002)	NA	NA	3
<i>Castanea</i> RAD-seq & Hyb-Seq	Combined	NA	NA	NA	1574	850847	540.6	44746 (0.053)	12316 (0.014)	1	1	1
<i>Hamamelis</i> RAD-seq	RAD-seq	NA	NA	All genes	1595	167828	105.2	5942 (0.035)	3353 (0.02)	1	2	1
<i>Hamamelis</i> Hyb-Seq	consensus	exon	untrimmed	orthologs	342	218141	637.8391813	8931 (0.041)	2934 (0.013)	2	2	3
			well-trimmed	orthologs	342	208406	609.374269	5690 (0.027)	1734 (0.008)	2	1	1
		supercontig	untrimmed	orthologs	342	605949	1771.8	53943 (0.089)	14001 (0.023)	2	2	3
			well-trimmed	orthologs	342	565626	1653.9	26093 (0.046)	7021 (0.023)	1	1	1
	degenerated (BWA19)	exon	auto-trimmed	All genes	348	198021	569.0	6242 (0.032)	2052 (0.01)	1	2	3
				orthologs	259	138509	534.8	4045 (0.029)	1299 (0.009)	2	2	2
				paralogs	92	58720	638.3	2085 (0.036)	726 (0.012)	1	3	3
			well-trimmed	All genes	352	216554	615.2	4577 (0.021)	1393 (0.006)	2	3	3
				orthologs	255	148936	584.1	3186 (0.021)	983 (0.007)	2	2	2
				paralogs	97	67618	697.1	1391 (0.021)	410 (0.006)	2	1	3
		supercontig	untrimmed	supercontig only	164	499452	3045.4	42223 (0.085)	11022 (0.022)	1	1	1

			auto-trimmed	All genes	344	261188	759.3	10134 (0.039)	2812 (0.011)	2	2	1
				orthologs	265	195148	736.4	7137 (0.037)	2124 (0.011)	2	2	1
				paralogs	87	69489	798.7	3030 (0.044)	696 (0.01)	1	2	2
			well-trimmed	supercontig only	164	473010	2884.2	18524 (0.039)	5099 (0.011)	1	1	1
				All genes	352	582238	1654.1	20852 (0.036)	5815 (0.01)	1	1	1
				orthologs	246	367117	1492.3	13263 (0.036)	3885 (0.011)	1	1	1
				paralogs	106	215121	2029.4	7589 (0.035)	1930 (0.009)	1	1	1
			well-trimmed	supercontig only (orthologs)	147	418666	2848.1	17745 (0.042)	4919 (0.012)	1	1	1
				supercontig only (paralogs)	13	42092	3237.8	1679 (0.04)	534 (0.013)	3	1	2
				All genes	346	568476	1643.0	22130 (0.039)	6292 (0.011)	1	1	1
				orthologs	319	514351	1612.4	20185 (0.039)	5693 (0.011)	1	1	1
				paralogs	27	54125	2004.6	1945 (0.036)	599 (0.011)	3	1	1
Hamamelis Hyb-Seq Chloroplast	Chloroplast (BWA100)	NA	well-trimmed	NA	1	107068	107068.0	589 (0.006)	70 (0.001)	NA	NA	2
Hamamelis RAD-seq & Hyb-Seq	Combined	NA	NA	NA	1912	678780	355.0	24823 (0.037)	6971 (0.010)	1	1	1

Notes: The Angiosperm 353 data have a set of matrices with different comparisons, include certain *consensus* vs. *degenerated*, exon vs. supercontig, orthologs vs. paralogs, and matrices with different trimming parameters (untrimmed, “auto-trimmed” trimming in Trimal, and our “well-trimmed” method in our PPD Pipeline) and mapping parameters (BWA -k). “*consensus*” matrix contains sequences with heterozygous sites represented by the most common code; “*degenerated*” matrix contains sequences with heterozygous sites represented by the IUPAC ambiguity codes. Numbers in species trees and gene trees indicate different topologies. In *Castanea*, three major topologies are included. Topology 1 indicates the topology with ((EA), (ENA, Europe)), topology 2 indicates the topology with ((EA, Europe), (ENA)), and topology 3 indicates the (*C. crenata*, (Europe, ENA, EA)). In *Hamamelis*, three major topologies are included. Topology 1 indicates the topology with ((*H. mollis*, (*H. japonica*, ENA)), topology 2 indicates (*H. mollis*, *H. japonica*), ENA)), and topology 3 indicates (*H. japonica*, (*H. mollis*, ENA)). MSA represents the multiple sequence alignment.

A NEW PARALOG REMOVAL PIPELINE

Phylogenetic Analyses

Phylogeny of RAD-seq data.—Preliminary phylogenetic analyses of all matrices (M20-M80) from RAD-seq were initially conducted using a partitioned maximum likelihood (ML) analysis implemented in the IQ-Tree program (Nguyen et al. 2015). All analyses used the TESTNEW option to obtain the best molecular model per partition. The results showed that the trees derived from different datasets were congruent in their topologies while the M50-based tree had the highest support values at most nodes. Due to this reason and the higher number of loci compared to M60-M80 and lower level of missing data compared to M20-M40 (see Results), we used the M50 dataset for final phylogenetic and downstream analyses. The M50 dataset of each genus was further filtered by removing samples containing 30% or more uncertain bases (Ns) and hybrid species *Hamamelis* × *intermedia* that contains many degenerate bases in the concatenated sequence. This filtering resulted in the removal of 2 *C. sativa* samples (CS_57 and CS_58) from *Castanea* and 3 samples (HI_85, HME_78, and HVE_75) from *Hamamelis*. A total of 25 samples of *Castanea* (plus one outgroup) and 22 samples of *Hamamelis* (plus three outgroups) representing all species of the two genera were remained for further analyses.

We conducted phylogenetic analyses of the filtered data using both the ML and Bayesian Inference (BI) methods. The ML analyses were partitioned by locus and were conducted using both IQ-Tree (Nguyen et al. 2015) and RAxML v8.2.4 (Stamatakis 2014), while the BI analysis was conducted using MrBayes v.3.2.2 (Ronquist and Huelsenbeck 2003). Analyses with IQ-Tree used the modelfinder (flag -m TESTNEW) and 1000 replicates of ultrafast bootstrap. The TESTNEW option was applied to find the best molecular model for each locus. In the ML analysis with RAxML, jModelTest v.2.1.4 (Darriba et al. 2012) was used to determine the best substitution model based on AICs (GTR+Γ and HKY+Γ models were chosen for *Castanea* and *Hamamelis*,

respectively), with 1000 bootstrap replicates. For BI analysis, the HKY+ Γ model was chosen for both *Castanea* and *Hamamelis* according to the BIC in jModelTest v.2.1.4. Two independent runs of Markov Chain Monte Carlo (MCMC) analyses were executed for each genus; each run was improved by the Metropolis Coupling method, which includes one cold chain and three heated chains that started randomly in the parameter space. Ten million generations with tree sampling in every 1000 generations and a burn-in of the first 25% of sampled trees were set for each run. Convergence was assessed by evaluation of standard deviation of split frequencies between tree samples (which approached zero, <0.001), the effective sample size (ESS >200 in all cases), and by evaluating the plot of the likelihood scores from the resulting MrBayes trees in Tracer (Rambaut et al. 2018). The IQ-Tree, RAxML, and BI analyses were all conducted on the CIPRES Science Gateway Portal (Miller et al. 2010).

We constructed species trees using ASTRAL-III (Zhang et al. 2018) and SVDquartets (Chifman and Kubatko 2014). For the analyses with ASTRAL-III, we used IQ-Tree as above to generate gene trees for both genera as inputs and ran ASTRAL-III with the default parameters. For the analyses with SVDquartets, PAUP* v4.0a166 (Swofford 2003) was used to generate a total of 100,000 quartets with 100 bootstrap replicates and then the quartet assembly method QFM was used to produce a summary tree (Reaz et al. 2014), following Zhou et al. (2020). All gene trees and species trees were visualized and edited in FigTree v.1.4.2 (Rambaut 2012) and edited with ggtree [R] (Yu et al. 2018) and Adobe Illustrator 2020 (Adobe Systems, San Francisco, CA, USA).

Phylogeny of Hyb-Seq data.—Phylogenetic analyses of the Hyb-Seq data were also performed for the various data matrices listed in Table 2 with IQ-Tree (Nguyen et al. 2015) to generate concatenated gene trees, and with both ASTRAL-III (Zhang et al. 2018) and SVDquartets

A NEW PARALOG REMOVAL PIPELINE

(Chifman and Kubatko 2014) to generate the species trees, as described above for the RAD-seq data. Phylogenetic analyses of the plastid DNA data were performed with both IQ-Tree and RAxML v8.2.4 (Stamatakis 2014).

Phylogeny of the combined data.—The combined RAD-seq and Hyb-Seq data (primary Hyb-Seq-PPD, see Results) matrix (referred to as RAD-Hyb-Seq) was built from M50 and “degenerated” supercontig matrix of orthologous genes containing accessions with both RAD-seq and Hyb-Seq data available. We checked overlapping sequences between the two datasets by mapping RAD-seq data to the target reference genes using ipyrad and manually removed the redundant sequences (see Results below). Phylogenetic analyses of the combined data were also performed using IQ-Tree (Nguyen et al. 2015), ASTRAL-III (Zhang et al. 2018), and SVDquartets (Chifman and Kubatko 2014) to infer the “total evidence” gene phylogeny and species tree of both *Castanea* and *Hamamelis*.

Dating the Phylogenies

BEAST2 2.6.2 (Bouckaert et al. 2014) was employed to estimate the divergence times of lineages within each genus. BEAST2.6.2 can consider information at heterozygous sites in divergence time estimation. The divergence time analyses were conducted for both the RAD-seq data and the Hyb-Seq data. We reduced the RAD-seq data in sampling to include only the samples with Hyb-Seq data in the analyses so that the two types of data included identical samples (Table 1). The only exception was that RAD-seq data were missing for *Castanea sativa* due to the failure of herbarium samples in the RAD-seq experiment and lack of silica dried leaf samples at the time the experiment was conducted. Among the different Hyb-Seq data matrices, we performed

divergence time dating analyses for several matrices for each genus, including the “*degenerated*” and “*consensus*” supercontig matrices for orthologous genes, the “*degenerated*” exon matrix for orthologous genes, the “*degenerated*” supercontig matrix for paralogous genes, the combined RAD-Hyb-Seq data, and matrices using different trimming processes to allow comparisons of supercontig and exon and the effect of coding methods, paralogous genes, and trimming methods. The stem ages of *Castanea* and *Hamamelis* were constrained based on fossil evidence, as 66 to 72 Ma (lognormal) and 50 to 56 Ma (lognormal), respectively (for details, see Supplementary Information, available on Dryad).

For the RAD-seq data, divergence-time analyses were run under the HKY+ Γ model for both *Castanea* and *Hamamelis* according to the result of jModelTest, an uncorrelated lognormal relaxed clock (Drummond et al., 2006), and the birth–death process model (Stadler, 2010). We set the mean 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., 2014). All the other prior parameters were determined with empty runs without the sequence data. The empty runs were conducted to check the appropriates of prior settings; the posterior distribution of the divergence times of the constrained nodes produced from the empty runs are expected to cover or be similar to the constrained values with ESS over 200. The priors in the empty runs were then used in the real divergence time analyses with data. 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. The maximum credibility tree of mean heights was then constructed using TreeAnnotator after discarding 20% trees as burn-in. We compared

A NEW PARALOG REMOVAL PIPELINE

different Hyb-Seq matrices using the same fossil calibrations and same priors except substitution models which were decided by BIC values from jModelTest. For analyses with the Hyb-Seq data, the divergence time analyses were conducted in the same way as for the RAD-seq data, except using the GTR+ Γ molecular model for all orthologous gene matrices and HKY+ Γ molecular model for paralogous gene matrices. For the RAD-Hyb-Seq data, the GTR+ Γ model was chosen in the divergence time analysis. These models were chosen based on jModelTest.

Biogeographic Analyses

Total-evidence dating.—For our biogeographic analyses, we added fossils of *Castanea* and *Hamamelis* to the phylogeny to provide a more accurate evolutionary framework for the study. *Castanea* and *Hamamelis* have a number of fossils from the Paleogene and Neogene periods in the Northern Hemisphere (listed in Table 3 and see details in Supplementary Information, available on Dryad). We estimated divergence times of both genera with a total-evidence dating approach (Ronquist et al. 2012; Gavryushkina et al. 2017) in BEAST2 (Bouckaert et al. 2014) with a matrix that included both extant and fossil taxa and data from molecules for extant taxa and morphological characters for fossils and extant taxa. We were able to score seven morphological traits available in fossils of *Castanea* and six morphological traits available in fossils of *Hamamelis* (Supplementary Table S1, available on Dryad). The molecular matrix used in the analysis was the “degenerated” supercontig matrix of orthologous genes from Hyb-Seq, as this data set contains the most informative characters of orthologous genes and supported phylogenetic relationships consistent with the completely resolved species tree and the RAD-seq trees (see Results below). The molecular characters were recorded as missing in the fossil taxa. Following previous studies (Gavryushkina et al. 2017; Zhou et al. 2020), the fossilized birth death model (FBD) and relaxed

480 clock log normal were selected for these analyses. The GTR+ Γ was chosen for the molecular
 481 substitution model and the Lewis MK model (Lewis 2001) was chosen for the morphological data.
 482 A preliminary analysis with FBD showed unresolved placement of most fossil taxa with
 483 unexpected divergence times, similarly reported in a recent simulation study on tip-dating method
 484 (Luo et al. 2020), probably due to homoplasy of the few morphological characters available.
 485 Therefore, we modified the analysis to constrain the fossils with their most closely related clades
 486 or species based on paleobotany literature, with dates for fossils set as the value in between the
 487 age range of each fossil (Table 3). We also enforced the species tree backbone derived from the
 488 molecular data in the analysis to reduce the run time and ensure the recovery of relationships of
 489 extant species depicted in the species tree. We ran the analysis for 200 million generations with
 490 sampling of trees every 10,000 generations. Quality of the runs and parameter convergence were
 491 assessed using Tracer v.1.7.1. The maximum credibility tree of median heights was constructed
 492 using TreeAnnotator (Bouckaert et al. 2014) after discarding 20% trees as burn-in. As a
 493 comparison, we also constructed a dated total evidence tree using a matrix containing the combined
 494 data from RAD-seq and Hyb-Seq and the morphological characters using the same processes
 495 described above. The total evidence dating analysis was also repeated with the combined RAD-
 496 Hyb-Seq data.

497 **Table 3. Fossils reported for *Castanea* and *Hamamelis*.**

Fossil	Geographic Origin (abbreviation)	Age (Ma)	Affinity	Reference
<i>Castanea</i>				
<i>Castanea miomollissima</i> Hu and Chaney	Montana, US and Japan (AC)	Miocene (5.3-23)	<i>C. mollissima</i>	Becker 1969; Ishida 1970
<i>Castanea tanaii</i> Huzioka	Korea (A)	Miocene (5.3-23)	<i>C. crenata</i>	Huzioka 1972
<i>Castanea miocrenata</i> Tanai et Onoe	Japan (A)	late Miocene and Pliocene (2.6-11.6)	<i>C. crenata</i>	Ozaki 1991
<i>Castanea sativa</i>	Lower Rhine Basin, Germany (E)	Pliocene (2.6-5.3)	<i>C. sativa</i>	Van der Burgh 1983
<i>Castanea atavia</i> Unger	Germany (E)	upper and mid Miocene (11.6-23)	<i>C. sativa</i>	Mai 2001
<i>Castanea basidentata</i> R.N. Lakhanpal	Oregon, US (C)	late Oligocene (23-27.8)	<i>C. orientalis</i> (<i>C. dentata</i>)	Lakhanpal 1958
<i>Castanea dolichophylla</i> Cockerell (<i>Castanea orientalis</i> Chaney)	Montana, US (C)	Oligocene (23-33.9)	<i>C. seguinii</i> and <i>C. henryii</i>	Becker 1969
	Oregon, US (C)	late Oligocene (23-27.8)	<i>C. seguinii</i> and <i>C. henryii</i>	MacGinitie 1953
	Washington, US (C)	late Miocene (5.3-11.6)	<i>C. seguinii</i> and <i>C. henryii</i>	Hoffman 1932
<i>Castanea spokaneensis</i> (Knowlton) Chaney and Axelrod	Southwestern Montana, US (C)	Oligocene (23-33.9)	<i>Lithocarpus klamathensis</i> and <i>C. dentata</i>	Becker 1973
	California Western US (C)	Miocene (5.3-23)	<i>Lithocarpus klamathensis</i> and <i>C. dentata</i>	Becker 1973
	Southwestern Montana, US (C)		<i>Lithocarpus klamathensis</i> and <i>C. dentata</i>	
<i>Castanea fujiyamae</i> Tanai	Jilin, China (A)	early Miocene (13.8-23)	NA	Becker 1972
	Kitami region, Japan (A)	middle Eocene (37.8-47.8)	NA	Manchester et al. 2005
<i>Castanea intermedia</i> Lesquereux	Colorado, US (C)	early Oligocene (27.8-33.9)	NA	Tanai 1970, 1995
<i>Castanea castaneaefolia</i> (Unger) Knowlton	Alaska, Canada (C)	Paleocene (56-66)	NA	Brown 1962
	California (C)	Eocene (33.9-55.8)	NA	Hollick 1936
	western Tennessee, US (B)	Miocene (5.3-23)	NA	Hollick 1936
Castaneoideae		Paleocene/Eocene (55-66)	<i>Castanea</i> Root	Nixon and Crepet 1989
<i>Protogagacea allonensis</i> Herendeen, Crane et Drinnan	Georgia, US (B)	Campanian (late Cretaceous) (72.1-83.6)	<i>Castanea</i> Root	Herendeen et al. 1995
<i>Antiquacupula sulcata</i> Sims, Herendeen, and Crane	Georgia, US (B)	Santonian (late Cretaceous) (83.6-86.3)	<i>Castanea</i> Root	Sims et al. 1998
<i>Hamamelis</i>				
<i>Hamamelis clarus</i> Hollick	Alaska (C)	Eocene (33.9-55.8)	NA	Hollick 1936
<i>Hamamelis kushiroensis</i> Tanai	Japan (A)	Oligocene (23-33.9)	<i>H. japonica</i>	Tanai 1970
<i>Hamamelis protojaponica</i> Tanai et N. Suzuki	Korea (A)	Miocene (5.3-23)	<i>H. japonica</i>	Huzioka 1972
<i>Hamamelites voyana</i> (Lesquereux) MacGinitie	Central Sierra Nevada (C)	middle Eocene (37.8-47.8)	<i>H. virginiana</i> and <i>H. mollis</i>	MacGinitie 1941
	Washinton (C)	Eocene (33.9-55.8)	NA	Burnham 1994

	<i>Hamawilsonia Boglei</i> Benedict, Pigg & DeVore gen. et sp. nov. (Hamamelidaceae)	North Dakota, US (C)	late Paleocene (56-59)	<i>Hamamelis</i> and <i>Sinowilsonia</i>	Benedict et al. 2008 Radtke et al. 2005; Xie et al. 2010
	<i>Fothergilla malloryi</i> Radtke, Pigg & Wehr	Washington, US (C)	early Eocene (49-50)	<i>Hamamelis</i> Root	
498	Notes: A = eastern Asia, B = eastern North America and Mexico, C = western North America, E = Europe. The affinities were				
499	assigned based on the description in the literatures. NA means data are not available.				

A NEW PARALOG REMOVAL PIPELINE

500 *Reconstructing the biogeographic history with DEC.*—We conducted the biogeographic analyses
501 using the Dispersal-Extinction-Cladogenesis (DEC) method (Ree and Smith 2008) based on the
502 dated total evidence trees (from Hyb-Seq data only and the combined RAD-Hyb-Seq data)
503 including all fossil species for *Castanea* and *Hamamelis*. The selection of this method was based
504 on a test in BioGeoBEARS (Matzke 2013) in RASP (Yu et al. 2015) that compared among the
505 Dispersal-Vicariance Analysis (DIVA; Ronquist and Cannatella 1997), Dispersal-Extinction-
506 Cladogenesis (DEC), and BayArea methods (Landis et al. 2013). We did not consider the
507 parameter *j* (“jump”) that represents the founder-event, allowing daughter species to “jump”
508 outside the geographical range of parental species (Matzke 2013) in the test, due to concerns raised
509 in Ree and Sanmartín (2018). Among the three methods compared, the DEC method had the
510 highest AICc value. We defined five geographic regions for both *Castanea* and *Hamamelis*, in the
511 DEC analyses to cover the distributions of modern and extinct taxa: EA (“A”); ENA and Mexico
512 (“B”); western North America (WNA) (“C”); and Europe (“E”) (Table 3). We combined Mexico
513 with eastern North America according to the more or less continuous distribution between *H.*
514 *mexicana* and ENA clade. We implemented a matrix of relative dispersal rates in time slices
515 between areas (Supplementary Table S2, available on Dryad) based on largely the physical
516 connections between areas, according to (Graham 1999a, 1999b, 2018; Tiffney and Manchester
517 2001; Mann 2007). The rates were arbitrarily set using the following criteria: as 0.001 for the time
518 slices when the two areas were physically disconnected or isolated by barriers (such as water or
519 dry areas); as 1.0 when the two areas were physically and ecologically connected, and as values
520 between 0.001 – 1.0 (e.g., 0.2, 0.5, or 0.75) when the two areas were partially connected by islands
521 that were separated at different levels (Supplementary Table S2, available on Dryad). For ancestral
522 ranges with disjunctions, e.g., AB, CE in both *Castanea* and *Hamamelis*, we also set the probability

value to 0.001 for low probability because such an ancestral distribution would require frequent intercontinental long distance dispersal to maintain the disjunct populations as a single gene pool of the ancestor across disjunct areas which we believe is unlikely for *Castanea* and *Hamamelis* given the expected short-distance dispersal of their seeds (heavy in *Castanea* and likely dispersed by mammals; gravity-dispersed in *Hamamelis* ejected by capsules). For both analyses, we implemented a constraint of maximum areas of four to allow all four areas of the Northern Hemisphere to be included in an ancestral range.

Test for ILS and Introgression

Both incomplete lineage sorting (ILS) and introgression can result in phylogenetic incongruences among genes. We applied the D-statistic (ABBA-BABA) method in Dsuite (Malinsky 2019) and phylogenetic network method to test which of these causes may explain the conflicts among gene trees in *Castanea* and *Hamamelis*. The D-statistic test provides a measure of introgression between taxa given a guide tree topology (((P1, P2), P3), O), where O represents a closely related outgroup taxon and P1 and P2 are taxa tested for signatures of introgression with P3. The null hypothesis is as follows: if ILS is the only source of gene tree discordance, then the frequency of derived alleles exclusively present in P1 and P3 is expected to be equal to the frequency of derived alleles exclusively present in P2 and P3 ($D = 0$). However, if the introgression happened between P3 and P2 or P3 and P1, it will result in an excess of shared derived alleles between the taxa that experienced introgression ($D > 0$ or $D < 0$). The Dsuite is a tool for calculating D-statistics based on VCF files, assessing the significance via the jackknife test, and generating two-tailed P-values based on the Z-scores. To explain conflicts between the concatenated-gene tree and species tree, we performed two D-statistic tests for *Castanea* and

A NEW PARALOG REMOVAL PIPELINE

Hamamelis to assess ILS vs. introgression among specific lineages involved in the phylogenetic conflicts. Specifically, for *Castanea*, we performed the test with the Hyb-Seq data (primary Hyb-Seq-PPD, see Results) (as it contains complete species samples while the RAD-seq data had *C. sativa* missing) and combined RAD-Hyb-Seq data. We tested for introgression events among ENA *Castanea*, EA *Castanea*, and European *Castanea* (*C. sativa*) due to the weak support for the sister relationship between *C. sativa* and the ENA clade in the species tree, but strong support in the concatenated gene tree (see Results) and among *C. crenata*, the clade consisting of the remaining eastern Asian species, and the clade containing the ENA-*C. sativa* due to their conflicting relationships between the chloroplast gene tree and the nuclear gene tree (see Results). For all tests, the outgroup species used in phylogenetic studies, *Quercus castaneifolia*, was used as the outgroup. For *Hamamelis*, due to the low support in both the concatenated gene tree and the species tree at the two lower nodes of the phylogeny, involving relationships among the *H. mollis*, *H. japonica* (both EA), and ENA *Hamamelis* clade, see Results), we conducted the test with three data sets, the RAD-seq, Hyb-Seq data, and combined RAD-Hyb-Seq data for the presence of introgression. Similarly, the same outgroup taxa used in phylogenetic studies were used as the outgroup for the D-statistic test. We also performed the test among *H. mexicana*, *H. vernalis*, and *H. ovalis* within the ENA clade, due to low support of their relationships, using *H. virginiana* as the outgroup. Input VCF files were generated with ipyrad for the RAD-seq data and with msa2vcf in Jvarkit (Lindenbaum 2015) for the Hyb-Seq data.

We further conducted analyses of phylogenetic networks to confirm the potential introgression events among species or lineages within each genus suggested by the D-statistic tests. For both *Castanea* and *Hamamelis*, we used the quartet-based maximum-pseudolikelihood approach implemented in PhyloNet v.3.6.4 (Than et al. 2008; Yu and Nakhleh 2015) for the

purpose. The set of rooted gene trees from the Hyb-Seq data (primary Hyb-Seq-PPD, see Results) were used as the input for PhyloNet. We did not use RAD-seq loci as their average length was short relative to Hyb-Seq data, resulting in unreliable gene-tree topologies. The number of runs of the search (-x) was set as 100 and the maximum number of reticulations was set as 2 according to the results from the D-statistic tests. All other parameters were left as default settings, as recommended by the authors.

RESULTS

RAD-seq and Hyb-Seq Data

Locus recovery.—From RAD-seq, an average of 659,529 and 698,799 reads per sample were generated in *Castanea* and *Hamamelis*, respectively. After quality filtering by the ipyrad pipeline, an average of 659,405 reads were retained in *Castanea* and 698,651 reads remained in *Hamamelis* for the subsequent cluster procedure (Supplementary Table S3, available on Dryad). The clustering process resulted in matrices of M20 to M80 containing 3597 to 415 loci, respectively, in *Castanea* and 3935 to 395 loci, respectively, in *Hamamelis* (Supplementary Table S4, available on Dryad). After filtering, the length of each locus from RAD-seq is approximately 105 bp for both *Castanea* and *Hamamelis* (Table 2).

For Hyb-Seq datasets, the number of loci varied depending on gene region, filtering strategy, and categorization of genes in the alignment (e.g. ortholog vs. paralog or exon vs. supercontig). For complete the loci recovery and sequence length information in different datasets, see Table 2. The average length per locus in exon region of well-trimmed “*degenerated*” matrix of orthologs was approximately 594 bp in *Castanea* and 615 bp in *Hamamelis*, while the average length per locus in the well-trimmed “*degenerated*” supercontig matrix was much longer,

A NEW PARALOG REMOVAL PIPELINE

approximately 2448 bp in *Castanea* and 1643 bp in *Hamamelis*. The information on loci number, alignment length, average length per locus, number of segregating sites, and number of parsimony informative sites of all matrices derived from different trimming methods are available in Table 2. Overall, many more loci were obtained from RAD-seq compared to Hyb-Seq (344 genes) but the average length per locus is much longer in the Hyb-Seq data, resulting in longer alignment lengths in Hyb-Seq matrices (Table 2).

For the combined RAD-Hyb-Seq data, we first checked for duplication of loci present in both data sets. We found 12 loci (out of 1290 in *Castanea*) and 2 loci (out of 1595 in *Hamamelis*) in the RAD-seq data of the two genera that could be mapped to the Hyb-Seq data. After removing these loci from the RAD-seq M50, we obtained 1574 loci for *Castanea* and 1912 loci for *Hamamelis* (Table 2) for the combined RAD-Hyb-Seq data. The chloroplast DNA data matrices extracted from the Hyb-Seq data due to off-targeting contained 12,647 bp for *Castanea* and 107,068 bp for *Hamamelis* (Table 2).

Paralog detection in Hyb-Seq data.—The gene matrices generated by HybPiper showed 11 putative paralogs (Gene 6048, 6954, 4951, 4724, 5940, 6387, 6570, 7583, 7324, 5138, 5941) out of a total of 344 genes in *Castanea*, but only two putative paralogs (Gene 5463, 5347) out of 344 genes in *Hamamelis*. In contrast, our PPD pipeline detected all paralogs detected by HybPiper, as well as many additional paralogs. For *Castanea*, the PPD detected a total of 48 paralogs, and for *Hamamelis*, PPD detected a total of 27 paralogs (Table 2).

Final Hyb-Seq datasets.—Although we completed numerous analyses on all data matrices presented in Table 2, we have chosen to present results in the main text primarily for the well-

trimmed “*degenerated* BWA100” supercontig matrix of orthologs generated from our PPD pipeline. We will refer to the *Castanea* and *Hamamelis* alignments generated through this pipeline as primary Hyb-Seq-PPD datasets. Results of additional analyses performed on alternative matrices were presented in Supplementary Information (available on Dryad).

Phylogenetic Analyses

Castanea - RAD-seq.—The analyses of all different matrices (M20, M30, ..., M80) with IQ-Tree yielded phylogenetic trees with congruent topology (Supplementary Figs. S1-S7, available on Dryad). The M50 matrix was selected for detailed phylogenetic and divergence time analyses (due to its relatively optimal levels of sampling on loci and taxa, and robust nodal support on the tree), which contains 1290 loci, 22 samples with less than 30% missing data, and 8500 SNPs (Supplementary Table S4, available on Dryad).

The concatenated gene trees from both ML and BI methods showed a split of the EA and ENA species in two monophyletic sister clades with full support (Fig. 3a (1) and Supplementary Figs. S8a and S9a available on Dryad). Within the EA clade, *C. henryi* diverges off first, followed by *C. mollissima*, which is sister to the *C. crenata*-*C. seguinii* clade. Within the ENA clade, a sister relationship between *C. pumila* and *C. dentata* is strongly supported. The species trees from both ASTRAL-III and SVDquartets based showed the same relationships with strong support (Fig. 3c).

A NEW PARALOG REMOVAL PIPELINE

Fig. 3. Gene trees and species tree of *Castanea* derived from different data sets. a) (1): the concatenated gene tree from RAD-seq M50 dataset; (2): the concatenated gene tree from well-trimmed “BWA100 *degenerated*” supercontig of orthologous gene data from Angiosperm 353 Hyb-Seq (primary Hyb-Seq-PPD data). Dash lines in the middle connecting the same samples used in both RAD-seq and Hyb-Seq analyses. Both concatenated gene trees in A were derived from analyses using the IQ-TREE whose topology is identical to those from RAXML and IQ-Tree. Numbers on branches are values of UF bootstrap from the IQ-TREE (only >90% numbers were shown), bootstrap values from RAXML (only >50% numbers were shown) (Supplementary Fig. S8, available on Dryad), and posterior probabilities from MrBayes (Supplementary Fig. S9,

available on Dryad), respectively. Asterisks represent 100% support values. b) ML trees from 12,647 bp of plastid gene captured in the Hyb-Seq data by RAxML and IQ-Tree. Numbers on branches are combined from bootstrap in RAxML and ultrafast bootstrap in IQ-Tree. c) The combined species tree from RAD-seq and Hyb-Seq data derived from analyses using ASTRAL-III and SVDQuartets. The numbers on the branches are support values from the RAD-seq data with ASTRAL-III, RAD-seq data with SVDQuartets, Hyb-Seq data with ASTRAL-III, and Hyb-Seq data with SVDQuartets, respectively. *Quercus castaneifolia* was used as the outgroup in all analysis. All trees were drawn by ggtree in R.

Castanea - Hyb-Seq.—We were able to obtain Hyb-Seq data for one specimen of *C. sativa* so that all species of the genus were represented. The phylogenetic trees from ML and BI analyses of the primary Hyb-Seq-PPD dataset (see above) revealed the same relationships as shown in the RAD-seq trees, recognizing the reciprocal monophyly of the EA and ENA species, with high support (Fig. 3a (2) and Supplementary Figs. S8b and S9b available on Dryad). The European species *C. sativa* is placed as the sister to the American clade, with high support (Fig. 3a (2) and Supplementary Figs. S8b and S9b available on Dryad).

The species trees constructed from this data matrix using ASTRAL-III and SVDQuartets have identical topologies that depict the same species relationships as the concatenated gene tree, with strong support for all nodes except one. The node uniting *C. sativa* and the ENA species in the species tree from SVDQuartets is relatively weakly supported (72% - see Figure 3c).

The analysis of the combined RAD-seq and primary Hyb-Seq-PPD data (the RAD-Hyb-Seq data) matrix resulted in identical ML and species tree topologies, which were also identical to the results of both RAD-seq and the primary Hyb-Seq-PPD datasets (see above). The combined

A NEW PARALOG REMOVAL PIPELINE

data based-species tree is also highly supported except for the node uniting *C. sativa* and the ENA clade (SVDQuartets 68%) (Supplementary Fig. S10, available on Dryad).

Castanea - chloroplast phylogeny.—Analyses of the chloroplast data using RAxML and IQ-tree resulted in an incompletely resolved phylogeny with lower support (Fig. 3b). The trees placed one of the EA species *C. crenata* as the sister of a well-supported clade consisting of the remaining species (93% bootstrap value and 95% ultrafast bootstrap value), different from the nuclear data from the RAD-seq and Hyb-Seq (Figs. 3a and 3c). Within this large clade, the remaining EA species form a well-supported monophyletic group (94% bootstrap value and 94% ultrafast bootstrap value). However, the relationships among this EA subclade, *C. sativa* from Europe, and the ENA species (*C. pumila* and *C. dentata*) remain unresolved (Fig. 3b).

Hamamelis - RAD-seq.—Similarly, we found little difference in tree topologies among the analyses of different matrices (M20, M30, ..., M80) using IQ-Tree. These trees varied slightly in support values (most pronounced in the tree from M80) (Supplementary Figs. S11-S17, available on Dryad). The M50 matrix was chosen for the same reasons described above for *Castanea* for downstream analyses (Supplementary Table S4, available on Dryad). Analyses of the filtered M50 with RAxML and MrBayes resulted in well-resolved phylogenies of the genus (Fig. 4a (1) and Supplementary Figs. S18a and S19a, available on Dryad). In trees from all three methods, one of the two EA species - *H. mollis* - is shown as the sister to all other *Hamamelis*, and the other EA species - *H. japonica* - is strongly supported as the sister of the ENA clade in the RAxML, IQ-Tree and MrBayes trees (with rapid bootstrap value 100, ultrafast bootstrap value 100 and PP value 1, respectively). The monophyly of the ENA clade is strongly supported, within which *H. ovalis* is

sister to *H. mexicana*. Species tree analyses resulted in similar topologies to the likelihood analyses, but with two notable exceptions. The species tree from ASTRAL showed a different sister-group relationship within the ENA clade (*H. ovalis* sister to *H. vernalis* – see Figure 4c), while the SVDQuartets analysis showed the two EA species as sister to one another, different from all other analyses of RAD-seq data performed, albeit with low support (Supplementary Fig. S20, available on Dryad).

Fig. 4. Gene trees and species tree of *Hamamelis* derived from different data sets. a) (1): the concatenated gene tree from RAD-seq M50 data; (2): the concatenated gene tree from the primary Hyb-Seq-PPD data. Dash lines in the middle connecting the same samples used in both

A NEW PARALOG REMOVAL PIPELINE

RAD-seq and Hyb-Seq approaches. Both concatenated gene trees in A were derived from analyses using the IQ-TREE, whose topology is identical to those from RAxML and IQ-Tree. Numbers on branches are values of UF bootstrap from IQ-Tree (only >90% numbers were shown), bootstrap values from RAxML (only >50% numbers were shown) (Supplementary Fig. S18, available on Dryad), and posterior probabilities from MrBayes (Supplementary Fig. S19, available on Dryad), respectively. Asterisks represent 100% support values. b) ML tree from 107,068 bp of plastid gene from Hyb-Seq data derived from analyses using RAxML and IQ-Tree. Numbers on branches are bootstrap values from analyses using RAxML and IQ-Tree, respectively. c) The combined species tree from RAD-seq and Hyb-Seq data derived from analyses using ASTRAL and SVDQuartets. Numbers on the branches are support values from ASTRAL-III (RAD-seq), ASTRAL-III (Hyb-Seq), and SVDQuartets (Hyb-Seq), respectively. The species tree from RAD-seq data with SVDQuartets method shows a different species tree and is presented in Supplementary Fig. S20, available on Dryad. *Parrotiopsis jacquemontiana* and *Fothergilla gardenii* were used as outgroups. All trees were drawn by ggtree in R.

Hamamelis - Hyb-Seq—The primary Hyb-Seq-PPD data for *Hamamelis* contains only 160 genes. We increased the gene sampling to 319 genes by adding the additional genes with exon data. The new primary data consisted of 514,351 bp with 5,693 sites that are parsimony informative (Table 2). The trees resulting from analyses of this matrix using IQ-Tree, RAxML, and MrBayes are identical in topology (Fig. 4a (2) and Supplementary Figs. S18b and S19b, available on Dryad) and show species relationships identical to those resolved by the concatenated RAD-seq data (Fig. 4a (1)), with strong support. The two species trees reconstructed from ASTRAL and SVDQuartets

are identical in topology (Fig. 4c), which is also identical to the concatenated Hyb-Seq gene tree and species tree inferred from the RAD-seq data with ASTRAL-III. Similarly, the node connecting the ENA clade and *H. japonica* is not well supported (0.59 in ASTRAL and 90 in SVDQuartets - see Figure 4c). The species trees inferred from the combined RAD-Hyb-Seq data were identical in topologies to trees derived from separate analyses of the RAD-seq and the primary Hyb-Seq-PPD dataset, but with higher support for the sister relationship between the ENA clade and *H. japonica* (0.99 in ASTRAL and 93 in SVDQuartets – see Supplementary Fig. S21, available on Dryad).

Hamamelis - chloroplast genes.—Analyses of the chloroplast DNA data from RAxML and IQ-tree recovered a well-supported phylogeny similar to the species tree from RAD-seq data with SVDQuartets that recognized the monophyly of the EA species, but different from the nuclear phylogeny from other analyses as mentioned above that supported *H. japonica* and the ENA clade being sister. Relationships among the four ENA species remain unresolved in the plastid gene tree (Fig. 4b).

Divergence Time Analyses

Castanea.—Divergence time analyses of the RAD-seq data estimated the crown age of the genus (splitting of the EA and ENA clades) as the early Miocene (23.2 Ma, 95% HPD: 21.4-25.8 Ma). Within the genus, other divergence occurred in the mid-Miocene and late Miocene (Fig. 5a). The divergences times estimated with the primary Hyb-Seq-PPD data are slightly younger (~0.6 to ~5.3 Ma) on the mean values for all of the nodes than those estimated from the RAD-seq data (Fig. 5c; Supplementary Table S5, available on Dryad). The European chestnut (*C. sativa*), which was not missing in the RAD-seq data, diverged from the two ENA species in the mid-Miocene (13.6

A NEW PARALOG REMOVAL PIPELINE

Ma, 95% HPD: 10.9-16.7 Ma) based on the primary Hyb-Seq-PPD data (Fig. 5c). The divergence times estimated from the combined RAD-Hyb-Seq data were slightly older for the mean values than those based on the primary Hyb-Seq-PPD data (Supplementary Fig. S22a, available on Dryad) and slightly younger than the values from the RAD-seq data.

Fig. 5 Results of divergence times estimated using RAD-seq and Hyb-Seq data. a) and b), results from RAD-seq M50 data; c) and d), results from the primary Hyb-Seq-PPD data. e) and f),

results from well-trimmed “*consensus*” Hyb-Seq supercontig data. All topology was drawn by
ggtree in R.

Hamamelis.—Analyses of the RAD-seq data, Hyb-Seq-PPD data, and combined RAD-Hyb-Seq data resulted in similar estimates of divergence times (Figs. 5b and 5d; Supplementary Fig. S22b, available on Dryad). The crown node of *Hamamelis* (splitting of *H. mollis* from the remaining species) was dated back to the late Oligocene (e.g., 25.9 Ma with the 95% HPD as 22.6-30.1 Ma from the RAD-seq data; Fig. 5b). The divergence of *H. japonica* from the ENA clade was dated to the early Miocene (e.g., 22.0 Ma, with the 95% HPD as 19.1-25.3 Ma from the RAD-seq data; Fig. 5b). Divergence events within the American clade were dated to the late Miocene for *H. virginiana* and the Pliocene for the other species (Fig. 5b).

Comparison of Alignments, Phylogenies and Divergence Times among Different Hyb-Seq Matrices

Phylogeny comparisons.—Five comparisons of the phylogeny inferred from different Hyb-Seq data sets were conducted to evaluate the impacts of paralogous genes and the value of our new pipeline (Table 2). The results showed that the “*consensus*” matrix from the HybPiper pipeline and the various “*degenerated*” matrices from our PPD pipeline resulted in trees with the same topology (Table 2; Supplementary Fig. S23, available on Dryad), but the trees derived from the “*consensus*” matrices had longer branch lengths than the trees derived from the “*degenerated*” matrices (Supplementary Fig. S23, available on Dryad), probably due to the greater number of parsimony informative sites in the “*consensus*” matrices (Table 2). The results also showed that the phylogenies inferred from the putative paralogs differed from the orthologous gene phylogenies in terms of topology and branch lengths (Table 2; Supplementary Fig. S24, available on Dryad). Our

A NEW PARALOG REMOVAL PIPELINE

results further showed congruence between the exon and supercontig trees in *Castanea* but minor incongruence between the exon and supercontig trees in *Hamamelis*. The exon tree grouped *H. mollis* and *H. japonica* as the sisters with low support (Supplementary Fig. S24, available on Dryad), a relationship not supported by analyses of the supercontig matrix (Figs. 4a (2) and 4c). Finally, our results showed that different levels of trimming of ortholog matrices did not influence topology, but did influence branch lengths (Supplementary Fig. S24, available on Dryad), whereas trimming paralogous gene alignments influenced both tree topologies and branch lengths (Supplementary Fig. S24 and details in Supplementary Information, available on Dryad).

Divergence time comparisons.—Divergence time analyses on three data matrices (RAD-seq data, primary Hyb-Seq-PPD data (“degenerated”), and “consensus” Hyb-Seq data) resulted in some notable difference between matrices, but with mean estimates that were largely within the range of the HPDs of other estimates. The divergence times estimated from the “consensus” Hyb-Seq data were slightly older (~2 - 4 Ma) at every node than those estimated from the primary Hyb-Seq-PPD data (Fig. 5; Supplementary Table S5, available on Dryad). The estimates based on the paralogs data identified by PPD were also older than those based on the orthologs and the 95% HPDs from the paralogs is also much larger in *Hamamelis* (Supplementary Fig. S25 and Fig. 5; Supplementary Table S5, available on Dryad). In contrast, the divergence times estimated from the exon and supercontig of orthologous genes are very similar at all nodes, with mean estimates overlapping HPDs of each analysis, although the exon based estimates are slightly older in *Castanea* and slightly younger in *Hamamelis*. Trimming the supercontig matrix influenced divergence times estimates to some extent; the “auto-trimmed” matrix produced older (2~6 Ma in *Castanea* and ~2 Ma in *Hamamelis*) estimates than the “well-trimmed” matrix (Supplementary

Fig. S25 and Fig. 5; Supplementary Table S5, available on Dryad), while the “untrimmed
consensus” matrix (i.e., data directly from HybPiper) produced much older estimates than the two
trimmed matrices (Supplementary Fig. S26 and Fig. 5; Supplementary Table S5, available on
Dryad; Table 2), although these matrices supported the same species relationship.

ILS and Introgression Tests

D-statistic tests.—Our analyses found significant signals rejecting the null ILS hypothesis ($D = 0$)
($P < 0.05$) in both *Castanea* and *Hamamelis*. The results suggested that, in *Castanea*, introgression
events occurred between the European species *C. sativa* (P2) and the EA clade (P3) (in both Hyb-
Seq and the combined RAD-Hyb-Seq data: $D = 0.3573$, $P < 0.0001$) and occurred between the EA
species *C. crenata* (P2) and ENA-*C. sativa* clade (P3) (Hyb-Seq: $D = 0.5123$, $P < 0.0001$; RAD-
Hyb-Seq: $D = 0.5135$, $P < 0.0001$) (Table 4). In *Hamamelis*, the analyses found significant signals
supporting introgression between *H. japonica* (as P3 or P2), and the American clade (as P2 or P3)
in the RAD-seq data ($P = 0.0015$), in the primary Hyb-Seq-PPD data ($P = 0.0009$), and in the
combined RAD-Hyb-Seq data ($P < 0.0001$; Table 4). However, the test did not find any significant
signals supporting introgression among *H. mexicana*, *H. vernalis*, and *H. ovalis* in these three data
sets (Table 4).

Table 4. D-statistic (ABBA-BABA) result for testing introgression events in *Castanea* and
Hamamelis.

Genus	Methods	O ^a	P1 ^a	P2 ^a	P3 ^a	D ^b	P
<i>Castanea</i>	Primary Hyb-Seq-PPD	<i>Quercus</i>	EA1	<i>C. crenata</i>	ENA and <i>C. sativa</i>	0.5123	<0.0001
		<i>Quercus</i>	ENA	<i>C. sativa</i>	EA	0.3573	<0.0001
	Combined RAD-Hyb-Seq	<i>Quercus</i>	EA1	<i>C. crenata</i>	ENA and <i>C. sativa</i>	0.5135	<0.0001
		<i>Quercus</i>	ENA	<i>C. sativa</i>	EA	0.3573	<0.0001
<i>Hamamelis</i>	RAD-seq	Outgroups of <i>Hamamelis</i>	<i>H. mollis</i>	ENA	<i>H. japonica</i>	0.3040	0.0015
		<i>H. virginiana</i>	<i>H. mexicana</i>	<i>H. vernalis</i>	<i>H. ovalis</i>	0.2381	0.1852

A NEW PARALOG REMOVAL PIPELINE

Primary Hyb-Seq-PPD	Outgroups of <i>Hamamelis</i>	<i>H. mollis</i>	<i>H. japonica</i>	ENA	0.1569	0.0009
	<i>H. virginiana</i>	<i>H. vernalis</i>	<i>H. ovalis</i>	<i>H. mexicana</i>	0.0680	0.1942
Combined RAD-Hyb-Seq	Outgroups of <i>Hamamelis</i>	<i>H. mollis</i>	<i>H. japonica</i>	ENA	0.1970	<0.0001
	<i>H. virginiana</i>	<i>H. ovalis</i>	<i>H. vernalis</i>	<i>H. mexicana</i>	0.0265	0.3569

Notes: ^a O represents a closely related outgroup taxon and P1 and P2 are taxa tested for signatures of introgression with P3. ^b when D significantly greater than 0, it indicates introgression events happened in P2 and P3. “ENA” means eastern North American clade, “EA” means eastern Asian clade, “EA1” means all eastern Asian species except *C. crenata*. “Primary Hyb-Seq-PPD” means the orthologous dataset generated by our pipeline. “Combined RAD-Hyb-Seq” means the dataset containing both RAD-seq and Primary Hyb-Seq-PPD data.

Phylogenetic Network.—The analyses with PhyloNet detected one reticulation event in *Castanea* between *C. sativa* and the stem of the EA clade excluding *C. henryi* (Fig. 6a). The inheritance probabilities show that *C. sativa* has 24% genomic contribution from the ancestor of EA clade excluding *C. henryi* (Fig. 6a). The analyses also detected one ancient reticulation event in *Hamamelis*, between *H. japonica* and the stem of the ENA clade. Inferred inheritance probabilities for this event indicate that the majority of the genomic contribution (65%) of the ENA clade came from its ancestor while a small proportion (35%) of the genomic contribution came from *H. japonica*. The tree topology from PhyloNet is similar to the chloroplast DNA tree showing the relationship of ((*H. mollis*, *H. japonica*), ENA clade) (Fig. 6b). The PhyloNet analyses did not detect the introgression event between *C. crenata* and ENA-*C. sativa* clade in *Castanea* suggested by the D-statistic test.

Fig. 6. Best supported species network of the orthologous genes from primary Hyb-Seq-PPD data inferred with PhyloNet. Numbers next to the hybrid branches indicate inheritance probabilities.

Biogeography of *Castanea* and *Hamamelis*

Castanea.—The biogeographic analyses with the DEC method based on the dated total evidence phylogenies resolved the ancestral area of *Castanea* as western North America (“C”) in the Paleocene, where the genus diversified and expanded the range into eastern Asia (“AC”) during the Eocene period (Fig. 7a - combined RAD-Hyb-Seq and morphology data; Supplementary Fig. S27a - primary Hyb-Seq-PPD data and morphology data, available on Dryad), then followed by further diversification in the two areas. During the Oligocene, the genus expanded the range into Europe and eastern North America (“ABCE”) and diversified again, leading to the origin of the common ancestor of the modern *Castanea* species in Eurasia and eastern North America (“ABE”) in the Miocene. The other lineages evolved prior to the Oligocene in eastern Asia and western North America are all represented by fossil species and their origins involved extinctions in Asia, Western North America, and Europe from the Paleocene through the Oligocene (Fig. 7a and Supplementary Fig. S27a, available on Dryad). A vicariance of the common ancestor of the modern

A NEW PARALOG REMOVAL PIPELINE

854 *Castanea* species in the early Miocene resulted in the EA (“A”) and Eur-ENA (“BE”) lineages.
855 The EA lineage later diversified further into at least seven species in the mid and late Miocene
856 with one of them spread into western North America. At least three of these species including the
857 one that expanded its range into western North America became extinct in the late Miocene (Fig.
858 7a). The lineage is now represented by the four EA species. The Eur-ENA lineage (“BE”) became
859 isolated between the two areas in the early-mid Miocene. The ENA part speciated into two extant
860 ENA species *C. pumila* and *C. dentata* in the mid-late Miocene, while the European (“E”) part
861 spread back to Asia and speciated further, followed by extinction in Asia in the early-mid Miocene.
862 One of them survived as *C. sativa* in Europe and the other is represented by a fossil species *C.*
863 *atavia* in Europe (Fig. 7a and Supplementary Fig. S27a, available on Dryad).

A NEW PARALOG REMOVAL PIPELINE

probabilities are shown by non-black colors in the pie charts. The black color in the pie chart indicates the total proportion (likelihood) of all other alternative ancestral ranges in the results of DEC analyses. Inferred intercontinental dispersals are indicated by red arrows. Fossil taxa are labeled in gray color. c) Distribution and hypothesized migration routes of *Castanea* including living species and fossils. d) Distribution and hypothesized migration routes of *Hamamelis* including living species and fossils. A: eastern Asia; B: eastern North America and Mexico; C: western North America; E: Europe.

Hamamelis.—The biogeographic analyses with the DEC method based on dated total evidence phylogeny of combined RAD-Hyb-Seq and morphology data resolved the ancestral area of *Hamamelis* as eastern Asia (“A”) in the mid-Eocene (47 Ma) (Fig. 7b - combined RAD-Hyb-Seq and morphology data; Supplementary Fig. S27b - primary Hyb-Seq-PPD data and morphology data, available on Dryad). The ancestor soon diverged into two lineages, one represented entirely by extinct fossil species, and the other containing both fossil and extant species. The lineage represented by fossil taxa spread into western North America (“C”) followed by subsequent diversification there, and dispersal back into Asia and Europe (“AC”). The other lineage spread to North America to obtain a widespread range in all three areas (“ABC”) before the Miocene and diverged into Asian and North American clades following the geographic isolation and extinction in western North America in the early Miocene (Fig. 7b and Supplementary Fig. S27b, available on Dryad). The Asian clade soon diversified into three lineages, with two surviving to present day. The North American clade diversified in the mid- and late-Miocene into four extant species of which one restricted to Mexico and the other three are restricted to ENA (“B”, Fig. 7b and Supplementary Fig. S27b, available on Dryad).

DISCUSSION

RAD-seq vs. Hyb-Seq Data

RAD-seq and targeted enrichment emerged to be two popular approaches in phylogenomic studies (e.g. Eaton and Ree 2013; Fernández-Mazuecos et al. 2018; Zhou et al. 2018; MacGuigan and Near 2019; Du et al. 2020 for RAD-seq data; Faircloth et al. 2013; McCormack et al. 2013; Leache et al. 2015; Lévillé-Bourret et al. 2018; Gaynor et al. 2020 for enrichment data), each with some notable advantages as discussed in the Introduction. We evaluated the two methods through side-by-side comparisons of phylogenomic results from data generated with RAD-seq and Hyb-Seq of Angiosperm 353 genes in the study of two EA-ENA disjunct genera, *Castanea* and *Hamamelis*. Our results showed, although there were fewer loci in the Angiosperm 353 gene data (e.g., 292 and 319 orthologs in well-trimmed “BWA100” datasets vs. 3,597 and 3,935 RAD-seq loci in the M20 matrices that contained the most RAD-seq loci of *Castanea* and *Hamamelis*, respectively) (Table 2 and Supplementary Table S4, available on Dryad), the total alignment length was much longer than that of the RAD-seq data owed to the longer average length of enrichment loci (Table 2). However, the proportion of the segregating sites in the Hyb-Seq data are similar to that of the RAD-seq data (Table 2). A similar pattern of differences in loci number, alignment length, and segregating sites between the RAD-seq and enrichment data was also reported in Harvey et al. (2016), but with lower values for these variables (the segregating sites: 4.07/site and 0.69% in the UCE matrices vs. 1.35/site and 1.41% in the RAD-seq matrices). However, the aforementioned differences in sequence data did not affect the topology recovered by phylogenetic methods in either genera (Figs. 3 and 4). Both of our RAD-seq and Hyb-Seq data well resolved the phylogenetic relationships within *Castanea* and *Hamamelis* with strong support (Figs. 3 and

A NEW PARALOG REMOVAL PIPELINE

4). This finding suggests that both approaches can generate a large amount of phylogenetic data for reliable inferences of species relationships within a genus. This is congruent with previous studies showing these two data sets performed similarly in phylogenetic inferences (e.g. Harvey et al. 2016; Manthey et al. 2016). However, the RAD-seq data may be unable to detect gene flow or introgression due to shortage of segregating sites per locus which decreases the chances of finding shared alleles on a very short phylogenetic branch (Harvey et al. 2016). We note that although both data supported the same species tree topology, the species tree may not be the correct species history, as found in our study of *Hamamelis* (see discussion below on introgression).

Although RAD-seq and Hyb-Seq data agreed in species relationships recovered from phylogenetic analyses, they appeared to slightly disagree in divergence times estimated using the same method. This result is unsurprising given the differences in alignment content and branch lengths in likelihood analyses. However, these differences had inconsistent effects across the two genera. The divergence times estimated from the Hyb-Seq data were younger (~1-5 Ma) in *Castanea* but older (~2-3 Ma) in *Hamamelis* at all nodes than those estimated from the RAD-seq data. Overall, the divergence time dating results are very similar between the RAD-seq and Hyb-Seq data (Fig. 5 and Supplementary Table S5, available on Dryad).

The phylogenetic congruence between data from RAD-seq and Hyb-Seq are encouraging and suggest that one may choose one or the other approach for a given phylogenomic study at the species level, according to their budget, labor and time available, conditions of DNA samples, and other goals. In general, the target enrichment approach can be effective with DNA samples at very low concentration or poor quality (Bi et al. 2012; Guschanski et al. 2013; McCormack et al. 2016), while the RAD-seq method requires DNA samples with relatively higher molecular weight. Although next-generation sequencing platforms have dramatically reduced the cost and time to

obtain genome-wide phylogenetic markers (Glenn 2011; Wetterstrand 2015), funding and time may still be a limitation in large scale comparative studies (Harris et al. 2010). In such concerns, RAD-seq would be a better choice over Hyb-Seq.

It is noteworthy that our RAD-seq data captured a small number of the Angiosperm genes included in the Hyb-Seq data (25 loci of RAD-seq matching to 21 loci of Hyb-Seq in *Castanea* and 13 loci of RAD-seq matching to 12 loci of Hyb-Seq in *Hamamelis*), indicating some level of overlap between the two datasets. This demonstrates that the RAD-seq data, as acquired using the experimental methods and analytical tools adopted in this study, likely represents a diverse array of genes and genomic regions, including the conserved low copy nuclear genes of angiosperms.

Impacts of Paralogs in Hyb-Seq Data from HybPiper, Sequence Coding Method, and the Value of the PPD

Our results also showed that our new pipeline (PPD) identified many more putative paralogs than HybPiper. Although the “consensus” sequence data generated from HybPiper produced 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), resulting in longer branches affecting divergence time estimation (Fig. 5 and Supplementary Fig. S26 and Table S5, 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. Our analyses of the loci containing paralogous genes from the Hyb-Seq data often resulted in phylogenies different from those inferred from data

A NEW PARALOG REMOVAL PIPELINE

of the orthologous genes (Supplementary Fig. S24, available on Dryad). The divergence times estimated from data including potential paralogous genes (i.e. the “*consensus*” data matrices from HypPiper) are likely older due to the additional variable sites introduced by gene paralogy (Supplementary Table S5, available on Dryad). Thus, our results clearly highlighted that paralogous gene content can negatively impact phylogenetic analyses. Our comparisons indicate that the PPD pipeline can effectively identify paralogs in the Hyb-Seq data and produce data with better quality for phylogenetic and divergence time dating analyses.

The advantage of “*degenerated*” (IUPAC consensus) coding of sequences from Hyb-Seq in phylogenetic studies has also been shown in previous studies. In Kates et al. (2018) the heterozygous sites were found to be disproportionately responsible for low resolution or topological uncertainty. In studies of Andermann et al. (2018b) and Lischer et al. (2014), although the IUPAC consensus matrix resulted in the same species tree as the “*consensus*” matrix, the divergence times estimated from the two data sets were different (underestimated divergence time in Andermann et al. 2018b and overestimated divergence time in Lischer et al. 2014). In our study, the “*consensus*” matrices resulted in long terminal branches (Supplementary Fig. S23, available on Dryad) and older divergence times at terminal nodes connecting species (dated to the mid-late Miocene in *Castanea* and the mid Miocene for the ENA species in *Hamamelis*) (Fig. 5). The longer branch lengths and older divergence times resulting from the “*consensus*” matrix (Fig. 5 and Supplementary Table S5, available on Dryad) are likely caused by mis-linkages of the most common SNPs within and among loci in the consensus matrix which resulted in inflated sequence variation among species.

In our PPD pipeline, we used the “*degenerated*” matrix using the IUPAC ambiguity codes instead of the allele phasing method because phasing many distant loci or phasing all heterozygous

positions within loci is still not possible. Most allele phasing methods only selected the two most common alleles as the two phasing, which might not truly reflect their actual phases (Lischer et al. 2014; Kates et al. 2018). In addition, many commonly used phylogenetic software packages have options to treat ambiguity-coded positions as informative, including RAxML (Stamatakis 2014), IQ-Tree (Nguyen et al. 2015) for tree building, SVDQuartets (Chifman and Kubatko 2014) for species tree, and BEAST2 (Bouckaert et al. 2014) for divergence time estimation. Therefore, our new PPD pipeline produces alignments suitable for analysis across a wide range of modern phylogenetic tools.

Influence of Gene Regions and Trimming Methods in Hyb-Seq Data

It is well-recognized that the rate and model of molecular evolution of genes and gene regions affect phylogenetic inferences. The sequence data from Hyb-Seq contain both exon and intron regions that evolve at strikingly different rates. How the exon and intron data would affect phylogenetic influences is worthy of exploration. Gene specific models of molecular evolution chosen from modeltest based on partitioned supercontig sequences may be less accurate, as they must account for the evolution of both the exon and intron regions in the same partition, which often evolve under different rates and models, and thus may not be accurate for either the exon or the intron regions within the same partition. The amount of phylogenetic information in the exon and intron regions also differs. How congruent are the analyses using the supercontig data with the approximate molecular models for the entire gene regions versus analyses using exon data alone with the more accurate models for exon regions remains to be explored. To gain insights into the question, we compared results from supercontig and exon data matrices constructed using the well-trimmed procedure and “degenerated” coding method.

A NEW PARALOG REMOVAL PIPELINE

Our results showed that in *Castanea*, the tree topologies inferred from the exon and supercontig data were identical, while in *Hamamelis*, the two datasets supported different relationships among *H. japonica*, *H. mollis*, and the ENA clade. The exon data supported a sister relationship between the two EA species *H. japonica* and *H. mollis*, with low support, consistent with the plastid data phylogeny where this relationship is strongly supported (Fig. 4 and Supplementary Fig. S24, available on Dryad), whereas the supercontig and RAD-seq data support a sister relationship between *H. japonica* and the ENA clade in all of the concatenated gene trees and species trees we inferred (Fig. 4). However, our PhyloNet result indicated that the relationships recovered by the exon data represented the true species history, while the relationship strongly supported by the supercontig data represents an ancient introgression (Fig. 6; see more details below). These results suggest that the information contained in exons can track a deeper true history, while the intron regions may track the more recent introgression history and be biased in species tree reconstruction, if the introgressed genomic contribution is high (35% in the case with *Hamamelis*; Fig. 6b). Our comparison between exon and supercontig data indicates the impact of rate of molecular evolution and suggests potential problems with the intron data from Hyb-seq and RAD-seq. We recommend vigorous exploration of datasets for more accurate reconstruction of phylogenetic histories of studied lineages.

Our comparisons among data sets trimmed differently by altering the matching length (k values in BWA implemented in our PPD pipeline) showed dramatic changes in the number of detected paralogs (48 vs. 116 in *Chastanea* and 27 vs. 106 in *Hamamelis* from k=19 to k=100 Table 2). Increasing the k value also resulted in the filtering of more mapping reads with low quality (matching length less than k). Therefore, we would like to suggest using our PPD pipeline

with larger k value (100) in BWA and “*degenerated*” coding to obtain Hyb-Seq data in better quality for phylogenetic analyses.

Biogeographic History of Castanea and Hamamelis and Insights into the EA-ENA Floristic Disjunction

The EA-ENA floristic disjunction represents a major phytogeographic pattern of in the Northern Hemisphere. Elements of the deciduous forests of the North Hemisphere exhibiting disjunct distributions in two or more areas (eastern Asia, eastern North America, western North America, and Europe) are often considered relics of the once widespread former mesophytic forests that continuously spanned the Northern Hemisphere during the late Paleogene and early Neogene (Wolfe 1980; Boufford and Spongberg 1983; Tiffney 1985a). The origin of the pattern is a major biogeographic puzzle of the earth biota which has attracted the attention of botanists for more than a century (Boufford and Spongberg 1983; Tiffney 1985a) and has been widely studied during the past 20 years (e.g., Wolfe 1980; Xiang et al. 1998a, 1998b, 2000; Wen 1999; Xiang and Soltis 2001; Donoghue and Smith 2004; Harris and Xiang 2009; Wen et al. 2010; Harris et al. 2013; Manos and Meireles 2015). Available evidence suggested the floristic disjunctions was the product of complex evolutionary processes including migration/dispersal, vicariance, extinction and speciation in different lineages at different times that were affected by geological and climatic changes in the past (Tiffney 1985a, 1985b; Wen 1999; Tiffney and Manchester 2001; Milne 2006; Manchester et al. 2009; Wen et al. 2010). However, the relative importance of these different processes in the formation of the phytogeographic pattern remains an open question and needs to be evaluated with evidence from individual disjunct taxa. Biogeographic studies of diverse disjunct genera integrating data from phylogeny, paleontology, and paleobotany are helpful to shed light

A NEW PARALOG REMOVAL PIPELINE

on the question. Previous studies supported that the Bering land bridge (BLB) and the North Atlantic land bridge (NALB) were two important corridors for plant exchanges between North America and Eurasia in the spread of the disjunct taxa during the Cenozoic period, with the NALB being more important for the thermophilic taxa (in the Paleogene) while the BLB being more important for the temperate taxa (in the Neogene) (Wen et al. 2010). Furthermore, mega analysis of biogeographic histories of ~100 lineages based on topologies of published phylogenies that mostly did not include fossils revealed bi-directional intercontinental dispersals, but “out of Asia” migration was more frequent (Donoghue and Smith 2004). These findings advanced our understanding of the biogeographic pattern but need to be updated with information from new studies employing better molecular data and more sophisticated analytical tools than what were available before. In our study of *Castanea* and *Hamamelis*, two of the many woody genera with living species isolated in EA and ENA (*Castanea* also in Europe) but many fossils in western North America (WNA) and Europe (Table 3), we set an example of such approach to provide new information for better understanding of the famous biogeographic pattern.

Our biogeographic analyses using the DEC method based on the dated total evidence tree suggests a western North American origin of *Castanea* in the Paleocene, with subsequent dispersals into Asia in the mid-Eocene and Europe-eastern North America in Oligocene (Fig. 7a). This result suggests a role of the BLB for the late Eocene dispersal into Asia and the role of the NALB for the late Oligocene dispersal into ENA and then into Europe. Paleontological evidence indicates that EA was physically connected to WNA by the BLB until ~4.7 Ma (Tiffney and Manchester 2001; Wen et al. 2016; Graham 2018), while NALB connected ENA and Europe in the early Paleogene (up to early Eocene). However, previous authors hypothesized that dispersal

and gene flow across the NALB may be possible after the early Eocene until the early Miocene via island hopping (Tiffney and Manchester 2001; Wen et al. 2016; Graham 2018). Our result also indicated that the most recent common ancestor of the extant species had a widespread range in EA, ENA, and Europe (ABE) in the early Miocene, supporting dispersal via NALB in the early Miocene. Our vigorous biogeographic analyses including fossils revealed a much more complicated history of the genus, involving Paleogene migration and extinctions prior to the origin of the present EA-ENA disjunct clade (Fig. 7a). Our study on *Castanea* also suggests that the present disjunct distribution of the genus in EA and ENA was a result of vicariance by geographic isolation and extinction and an “out of western North American” migration in the early history of the genus, in contrast to the major pattern of “out of Asia” migration detected in the previous mega-analysis (Donoghue and Smith 2004).

In *Hamamelis*, our study including fossils similarly revealed a more complicated biogeographic history of the genus than that of the extant taxa. An Asian distribution of the common ancestor of the genus in the early Eocene and a widespread ancestor of the extant lineage in Asia and North America in the late Oligocene supports “out of Asia” migration in the early history but a vicariance origin of the disjunct distribution of the modern taxa, respectively (Fig. 7b). In both genera, however, the divergence of the EA and ENA lineages occurred in the late Oligocene-Miocene time, agreeing with the major pattern found in previous studies (Donoghue and Smith 2004; Harris et al. 2013; reviewed in Wen et al. 2010). Our data support Paleogene dispersals via both the BLB and NALB and revealed many extinctions during the Eocene through Miocene in all four areas, especially more in EA and WNA in both genera (Fig. 7). Our data also indicate that *Castanea* and *Hamamelis* are relic taxa of the boreotropical flora evolved in the Paleocene and Eocene, respectively, and experienced a complicated biogeographical history in the

A NEW PARALOG REMOVAL PIPELINE

past, whereas the modern species are descendants of Mesophytic Forests developed after the global climatic cooling in the early Oligocene (Tiffney 1985b).

Evolutionary History and Ancient Introgression

In our study, both RAD-seq and Hyb-Seq data strongly supported the monophyly of each of *Castanea* and *Hamamelis*, and the monophyly of an EA species and the ENA species in each genus, respectively (Figs. 3 and 4). Both data also strongly supported the presence of ancient introgression between the European species *C. sativa* and the ancestor of EA clade after the divergence of *C. henryi* in *Castanea* and between *H. japonica* and the ancestor of the ENA clade of *Hamamelis* (Table 4, Figs. 6b and 7b). The ancestral ranges of *C. sativa* (“AE”) and EA clade (“A”) were overlapping in the Miocene and the introgression between the two lineages were possible. Morphologically, *C. sativa* does seem to show some intermediary between the EA and ENA clades (e.g. leaf size and pubescence on the abaxial surface of mature leaves). Our PhyloNet result revealed that the lineage leading to *Hamamelis japonica* contributed to 35% of the genome of the ENA clade (Fig. 6b). The age of *H. japonica* and its extinct ancestor fossil species (*H. protojaponica*) overlapped with the age of the ancestor of the ENA clade of *Hamamelis* in the mid-Miocene, lending evidence in agreement with the suggested introgression (Fig. 7b). Given the biogeographic history and ancestral range of the common ancestor of the two lineages (“ABC”), the divergence and subsequent introgression of these lineages likely occurred at the high latitudes of “ABC” along the BLB area where the lineages leading to *H. japonica* and the ENA clade were overlapping in range and hybridized before they became extinct in the BLB and western North America. A hybrid origin of *H. japonica* between the ancestor of *H. mollis* and the ancestor of the ENA lineage is also possible.

Ancient inter-clade or inter-species ingression have been also been uncovered in other lineages using the D-statistics method with enrichment data (Zarza et al. 2016; Alexander et al. 2017; Everson et al. 2019). However, the D-statistics method may be prone to false-positives due to model violations or outliers as could come from, e.g., undetected paralogy or even library construction artifacts (Lambert et al. 2019; Blair and Ané 2020). Our result similarly showed two introgression scenarios supported by D-statistics but not by PhyloNet (Table 4 and Fig. 6). Specifically, the D-statistic method suggested introgression between the eastern Asian *C. crenata* and the European *C. sativa*, and between *C. crenata* and the ENA clade, but these events were not revealed by PhyloNet (Fig. 6a). We suggest using D-statistics for an initial test, followed by analysis with PhyloNet for testing gene introgression for observed gene tree conflicts. PhyloNet infers phylogenetic networks with consideration of ILS (Yu et al. 2013) and is a more reliable approach than D-statistics alone.

In addition, our study of *Castanea* indicated that the plastid genome data extracted from the Hyb-Seq are valuable to reconstruct a more complete picture of the evolutionary history of a study group. By integrating evidence from the nuclear and plastid gene phylogenies we uncovered a chloroplast DNA capture event in *C. crenata* based on a strongly supported conflicting placement of the species between the plastid and nuclear gene trees (Fig. 3). Based on the plastid gene tree, the chloroplast DNA donor could be one of the extinct species of *Castanea*, such as *C. fujiyamae* that occurred in EA during 47.8 to 27.8 Ma (Fig. 6) or an outgroup taxon of the genus, such as *Castanopsis* which is closely related to *Castanea* and overlaps distribution with *C. crenata* in EA (Jisaburō et al. 1965).

Our phylogenetic data do not agree with the morphology-based classification scheme of three sections in *Castanea* (Sect. *Eucastanon*, Sect. *Balanocastanon*, and Sect. *Hypocastanon*)

A NEW PARALOG REMOVAL PIPELINE

1144 (Doye 1908; Johnson 1988). Our result indicated that Sect. *Eucastanon* that included *C. dentata*,
 1145 *C. sativa*, *C. mollissima*, *C. seguinii*, and *C. crenata* is paraphyletic and the character of one nut
 1146 per cupule in *C. pumila* (ENA) and *C. henryi* (EA) is homoplasious. The taxonomic status of the
 1147 Allegheny chinkapin (*C. pumila*) and the Ozark chinkapin has been disputed (Johnson 1988; Nixon
 1148 1997). Johnson (1988) considered the Ozark chinkapin as a variety of *C. pumila*, while Nixon
 1149 (1997) regarded it as a separate species *C. ozarkensis*. In our study, all individuals representing *C.*
 1150 *pumila* including the Ozark chinkapins formed a monophyletic group sister to *C. dentata* with
 1151 strong support. Therefore, our phylogenomic study do not support the recognition of *C. ozarkensis*
 1152 as a distinct species. However, the hypothesis should be further tested with population level
 1153 sampling of related taxa.

1154 It is noteworthy that the species relationship revealed from PhyloNet is congruent with the
 1155 chloroplast gene tree in grouping *H. japonica* and *H. mollis* as sisters (Fig. 4b). This relationship
 1156 was revealed in the exon orthologous data (Supplementary Fig. S24, available on Dryad) and one
 1157 of the species trees that was reconstructed by SVDquartets with the RAD-seq data (Supplementary
 1158 Fig. S20, available on Dryad). A simulation study by Chou et al (2015) showed that when ILS is
 1159 very high, ASTRAL performed better than SVDquartets in constructing species history, but when
 1160 ILS is low SVD quartet performed better than ASTRAL. In a more recent simulation study, Long
 1161 and Kubatko (2018) showed that when gene flow is present, ASTRAL is not consistent while
 1162 SVDquartets performed well. Our results appears to agree with the finding of Long and Kubatko
 1163 (2018) suggesting SVDquartets is more likely to recover the true species history with data from
 1164 numerous gene loci (like the RAD-seq data) when there is high level of inter-species introgression
 1165 as observed in *Hamamelis*. Nonetheless, in all species trees, the support for the sister taxon of the
 1166 ENA clade was low, clearly indicating conflicts among gene trees. Our results further indicate that

ZHOU ET AL.

in addition to species tree methods, data type also affects phylogenetic inferences in the presence of introgression.

SUPPLEMENTARY MATERIAL

Data available from the Dryad Digital Repository:

<https://doi.org/10.5061/dryad.2fqz612n3>

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, and a Shiu-Ying Hu Student/Post-Doctoral Exchange Award to W Zhou from the Arnold Arboretum. J. Soghigian was supported by NSF DEB – 1754376.

ACKNOWLEDGEMENTS

We thank the Soltis lab at Florida Museum of Natural History, Fu lab at Zhejiang University, Gao lab at Kunming Institute of Botany, Chinese Academy of Sciences, JC Raulston Arboretum, Arnold Arboretum, University of Washington Botanic Gardens, Paul Jones at Sarah Duke Garden, and UNC herbarium for providing some leaf tissues, Jeffrey Thorne from North Carolina State University and Matt Johnson from Texas Tech University for discussion of the PPD pipeline. We

A NEW PARALOG REMOVAL PIPELINE

1189 also thank Steven Manchester and Kathleen Pigg for the discussion of *Castanea* and *Hamamelis*
1190 fossils.

1191

1192 REFERENCES

1193 Alexander A.M., Su Y.-C., Oliveros C.H., Olson K.V., Travers S.L., Brown R.M. 2017. Genomic
1194 data reveals potential for hybridization, introgression, and incomplete lineage sorting to
1195 confound phylogenetic relationships in an adaptive radiation of narrow-mouth frogs.
1196 *Evolution*. 71:475–488.

1197 Altenhoff A.M., Glover N.M., Dessimoz C. 2019. Inferring orthology and paralogy. In
1198 *Evolutionary Genomics*. New York: Humana Press. p. 149–175.

1199 Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. 1990. Basic local alignment search
1200 tool. *J. Mol. Biol.* 215:403–410.

1201 Andermann T., Cano Á., Zizka A., Bacon C., Antonelli A. 2018a. SECAPR—a bioinformatics
1202 pipeline for the rapid and user-friendly processing of targeted enriched Illumina
1203 sequences, from raw reads to alignments. *PeerJ*. 6:e5175.

1204 Andermann T., Fernandes A.M., Olsson U., Töpel M., Pfeil B., Oxelman B., Aleixo A., Faircloth
1205 B.C., Antonelli A. 2018b. Allele Phasing Greatly Improves the Phylogenetic Utility of
1206 Ultraconserved Elements. *Syst. Biol.* 68(1):32–46.

1207 Andrews K.R., Good J.M., Miller M.R., Luikart G., Hohenlohe P.A. 2016. Harnessing the power
1208 of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.* 17:81–92.

1209 Arnold B., Corbett-Detig R.B., Hartl D., Bomblies K. 2013. RADseq underestimates diversity
1210 and introduces genealogical biases due to nonrandom haplotype sampling. *Mol. Ecol.*
1211 22:3179–3190.

1212 Bankevich A., Nurk S., Antipov D., Gurevich A.A., Dvorkin M., Kulikov A.S., Lesin V.M.,
1213 Nikolenko S.I., Pham S., Prjibelski A.D., Pyshkin A.V., Sirotkin A.V., Vyahhi N., Tesler
1214 G., Alekseyev M.A., Pevzner P.A. 2012. SPAdes: A New Genome Assembly Algorithm
1215 and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* 19:455–477.

1216 Becker H.F. 1969. Fossil plants of the Tertiary Beaverhead Basins in southwestern Montana.
1217 *Palaeontographica, Abt. B.* 127:1–142.

1218 Becker H.F. 1972. The Metzel Ranch flora of the upper Ruby River basin, southwestern
1219 Montana. *Palaeontographica Abt. B.* 141:1–61.

1220 Becker H.F. 1973. The York Ranch flora of the upper Ruby River basin, southwestern Montana.
1221 *Palaeontographica Abt. B.* 143:18–93.

- 1222 Benedict J.C., Pigg K.B., DeVore M.L. 2008. *Hamawilsonia boglei* gen. et sp.
1223 nov.(Hamamelidaceae) from the late Paleocene Almont flora of central North Dakota. Int.
1224 J. Plant Sci. 169:687–700.
- 1225 Bi K., Vanderpool D., Singhal S., Linderoth T., Moritz C., Good J.M. 2012. Transcriptome-based
1226 exon capture enables highly cost-effective comparative genomic data collection at
1227 moderate evolutionary scales. BMC Genom. 13:403.
- 1228 Blair C., Ané C. 2020. Phylogenetic Trees and Networks Can Serve as Powerful and
1229 Complementary Approaches for Analysis of Genomic Data. Syst. Biol. 69:593–601.
- 1230 Bolger A.M., Lohse M., Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence
1231 data. Bioinformatics. 30:2114–2120.
- 1232 Bouckaert R., Heled J., Kuhnert D., Vaughan T., Wu C.H., Xie D., Suchard M.A., Rambaut A.,
1233 Drummond A.J. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis.
1234 PLoS Comput. Biol. 10:e1003537.
- 1235 Boufford D.E., Spongberg S.A. 1983. Eastern Asian-Eastern North American Phytogeographical
1236 Relationships-A History From the Time of Linnaeus to the Twentieth Century. Ann. Mo.
1237 Bot. Gard. 70:423.
- 1238 Bradford J.L., Marsh D.L. 1977. Comparative Studies of the Witch Hazels *Hamamelis*
1239 *virginiana* and *H. vernalis*. J. Ark. Acad. Sci. 31(1): 29-31.
- 1240 Brown R.W. 1962. Paleocene flora of the Rocky Mountains and great Plains. Washington: U.S.
1241 Govt. Print. Off.p. 1-119.
- 1242 Burnham R.J. 1994. Paleoecological and floristic heterogeneity in the plant-fossil record: an
1243 analysis based on the Eocene of Washington. Washington: U.S. Govt. Print. Off.p. 1-31.
- 1244 Chanderbali A.S., Berger B.A., Howarth D.G., Soltis D.E., Soltis P.S. 2017. Evolution of floral
1245 diversity: genomics, genes and *gamma*. Phil. Trans. R. Soc. B. 372:20150509.
- 1246 Chang H.T. 1979. Hamamelidaceae. In: Florae Reipublicae Popularis Sinicae. Beijing: Science
1247 Press. 35(2): 36-116.
- 1248 Chifman J., Kubatko L. 2014. Quartet inference from SNP data under the coalescent model.
1249 Bioinformatics. 30:3317–3324.
- 1250 Cullings K.W. 1992. Design and testing of a plant-specific PCR primer for ecological and
1251 evolutionary studies. Mol. Ecol. 1:233–240.
- 1252 DaCosta J.M., Sorenson M.D. 2014. Amplification Biases and Consistent Recovery of Loci in a
1253 Double-Digest RAD-seq Protocol. PLoS ONE. 9:e106713.

A NEW PARALOG REMOVAL PIPELINE

- 1254 DaCosta J.M., Sorenson M.D. 2016. ddRAD-seq phylogenetics based on nucleotide, indel, and
1255 presence-absence polymorphisms: Analyses of two avian genera with contrasting
1256 histories. *Mol. Phylogenet. Evol.* 94:122–135.
- 1257 Darriba D., Taboada G.L., Doallo R., Posada D. 2012. jModelTest 2: more models, new
1258 heuristics and parallel computing. *Nat. Methods.* 9:772.
- 1259 DePristo M.A., Banks E., Poplin R., Garimella K.V., Maguire J.R., Hartl C., Philippakis A.A.,
1260 del Angel G., Rivas M.A., Hanna M., McKenna A., Fennell T.J., Kernysky A.M.,
1261 Sivachenko A.Y., Cibulskis K., Gabriel S.B., Altshuler D., Daly M.J. 2011. A framework
1262 for variation discovery and genotyping using next-generation DNA sequencing data. *Nat.*
1263 *Genet.* 43:491–498.
- 1264 Dode L.A. 1908. Notes dendrologiques. Paris: Au Siège de la Société.p. 1-166.
- 1265 Dodsworth S., Pokorny L., Johnson M.G., Kim J.T., Maurin O., Wickett N.J., Forest F., Baker
1266 W.J. 2019. Hyb-Seq for Flowering Plant Systematics. *Trends Plant Sci.* 24:887–891.
- 1267 Dong W., Xu C., Wu P., Cheng T., Yu J., Zhou S., Hong D.-Y. 2018. Resolving the systematic
1268 positions of enigmatic taxa: Manipulating the chloroplast genome data of Saxifragales.
1269 *Mol. Phylogenet. Evol.* 126:321–330.
- 1270 Dong Y., Chen S., Cheng S., Zhou W., Ma Q., Chen Z., Fu C.-X., Liu X., Zhao Y., Soltis P.S.,
1271 Wong G.K.-S., Soltis D.E., Xiang Q.-Y. 2019. Natural selection and repeated patterns of
1272 molecular evolution following allopatric divergence. *eLife.* 8:e45199.
- 1273 Donoghue M.J., Smith S.A. 2004. Patterns in the assembly of temperate forests around the
1274 Northern Hemisphere. *Phil. Trans. R. Soc. Lond. B.* 359:1633–1644.
- 1275 Doyle J. 1991. DNA protocols for plants. *Molecular techniques in taxonomy.* Springer-Verlag. p.
1276 283–293.
- 1277 Du Z.-Y., Harris A., Xiang Q.-Y. (Jenny). 2020. Phylogenomics, co-evolution of ecological niche
1278 and morphology, and historical biogeography of buckeyes, horsechestnuts, and their
1279 relatives (Hippocastaneae, Sapindaceae) and the value of RAD-Seq for deep evolutionary
1280 inferences back to the Late Cretaceous. *Mol. Phylogenet. Evol.* 145:106726.
- 1281 Eaton D.A., Ree R.H. 2013. Inferring phylogeny and introgression using RADseq data: an
1282 example from flowering plants (Pedicularis: Orobanchaceae). *Syst. Biol.* 62:689–706.
- 1283 Eaton D.A.R. 2014. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses.
1284 *Bioinformatics.* 30:1844–1849.
- 1285 Eaton D.A.R., Overcast I. 2020. ipyrad: Interactive assembly and analysis of RADseq datasets.
1286 *Bioinformatics.* 36:2592–2594.

- 1287 Eaton D.A.R., Spriggs E.L., Park B., Donoghue M.J. 2016. Misconceptions on Missing Data in
1288 RAD-seq Phylogenetics with a Deep-scale Example from Flowering Plants. *Syst. Biol.*
1289 66(3): 399–412.
- 1290 Edgar R.C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space
1291 complexity. *BMC bioinform.* 5:113.
- 1292 Everson K.M., McLaughlin J.F., Cato I.A., Evans M.M., Gastaldi A.R., Mills K.K., Shink K.G.,
1293 Wilbur S.M., Winker K. 2019. Speciation, gene flow, and seasonal migration in *Catharus*
1294 thrushes (Aves: Turdidae). *Mol. Phylogenet and Evol.* 139:106564.
- 1295 Faircloth B.C. 2016. PHYLUCE is a software package for the analysis of conserved genomic
1296 loci. *Bioinformatics.* 32:786–788.
- 1297 Faircloth B.C., Sorenson L., Santini F., Alfaro M.E. 2013. A Phylogenomic Perspective on the
1298 Radiation of Ray-Finned Fishes Based upon Targeted Sequencing of Ultraconserved
1299 Elements (UCEs). *PLoS ONE.* 8:e65923.
- 1300 Fernández-Mazuecos M., Mellers G., Vigalondo B., Sáez L., Vargas P., Glover B.J. 2018.
1301 Resolving Recent Plant Radiations: Power and Robustness of Genotyping-by-
1302 Sequencing. *Syst. Biol.* 67:250–268.
- 1303 Flanagan S.P., Jones A.G. 2018. Substantial differences in bias between single-digest and double-
1304 digest RAD-seq libraries: A case study. *Mol. Ecol. Resour.* 18:264–280.
- 1305 Funk V.A. 2018. Collections-based science in the 21st Century: Collections-based science in the
1306 21st Century. *J. Syst. Evol.* 56:175–193.
- 1307 Gautier M., Gharbi K., Cezard T., Foucaud J., Kerdelhué C., Pudlo P., Cornuet J.-M., Estoup A.
1308 2013. The effect of RAD allele dropout on the estimation of genetic variation within and
1309 between populations. *Mol. Ecol.* 22:3165–3178.
- 1310 Gavryushkina A., Heath T.A., Ksepka D.T., Stadler T., Welch D., Drummond A.J. 2017.
1311 Bayesian total-evidence dating reveals the recent crown radiation of penguins. *Syst. Biol.*
1312 66:57–73.
- 1313 Gaynor M.L., Fu C., Gao L., Lu L., Soltis D.E., Soltis P.S. 2020. Biogeography and ecological
1314 niche evolution in Diapensiaceae inferred from phylogenetic analysis. *J. Syst. Evol.*
1315 58(5): 646–662.
- 1316 Glenn T.C. 2011. Field guide to next-generation DNA sequencers: Field Guide to Next-Gen
1317 Sequencers. *Mol. Ecol. Resour.* 11:759–769.
- 1318 Graham A. 1999a. The Tertiary history of the northern temperate element in the northern Latin
1319 American biota. *Am. J. Bot.* 86:32–38.
- 1320 Graham A. 1999b. Late Cretaceous and Cenozoic history of North American vegetation: north of
1321 Mexico. New York: Oxford University Press.

A NEW PARALOG REMOVAL PIPELINE

- 1322 Graham A. 2018. The role of land bridges, ancient environments, and migrations in the assembly
1323 of the North American flora. *J. Syst. Evol.* 56:405–429.
- 1324 Guschanski K., Krause J., Sawyer S., Valente L.M., Bailey S., Finstermeier K., Sabin R.,
1325 Gilissen E., Sonet G., Nagy Z.T., Lenglet G., Mayer F., Savolainen V. 2013. Next-
1326 Generation Museomics Disentangles One of the Largest Primate Radiations. *Syst. Biol.*
1327 62:539–554.
- 1328 Harris R.S. 2007. Improved pairwise Alignment of genomic DNA. PhD thesis. Penn. State Univ.
- 1329 Harris A., Wen J., Xiang Q.-Y.J. 2013. Inferring the biogeographic origins of inter-continental
1330 disjunct endemics using a Bayes-DIVA approach: Biogeography of continental endemic
1331 species. *J. of Syst. Evol.* 51:117–133.
- 1332 Harris A.J., Xiang Q.Y. 2009. Estimating ancestral distributions of lineages with uncertain sister
1333 groups: A statistical approach to dispersal-vicariance analysis and a case using *Aesculus*
1334 L. (Sapindaceae) including fossils. *J. Syst. Evol.* 47:349–368.
- 1335 Harris R.A., Wang T., Coarfa C., Nagarajan R.P., Hong C., Downey S.L., Johnson B.E., Fouse
1336 S.D., Delaney A., Zhao Y., Olshen A., Ballinger T., Zhou X., Forsberg K.J., Gu J.,
1337 Echipare L., O’Geen H., Lister R., Pelizzola M., Xi Y., Epstein C.B., Bernstein B.E.,
1338 Hawkins R.D., Ren B., Chung W.-Y., Gu H., Bock C., Gnirke A., Zhang M.Q., Haussler
1339 D., Ecker J.R., Li W., Farnham P.J., Waterland R.A., Meissner A., Marra M.A., Hirst M.,
1340 Milosavljevic A., Costello J.F. 2010. Comparison of sequencing-based methods to profile
1341 DNA methylation and identification of monoallelic epigenetic modifications. *Nat.*
1342 *Biotechnol.* 28:1097–1105.
- 1343 Harvey M.G., Smith B.T., Glenn T.C., Faircloth B.C., Brumfield R.T. 2016. Sequence Capture
1344 versus Restriction Site Associated DNA Sequencing for Shallow Systematics. *Syst. Biol.*
1345 65:910–924.
- 1346 Herendeen P.S., Crane P.R., Drinnan A.N. 1995. Fagaceous flowers, fruits, and cupules from the
1347 Campanian (Late Cretaceous) of central Georgia, USA. *Int. J. Plant Sci.* 156:93–116.
- 1348 Herrando-Moraira S., Calleja J.A., Carnicero P., Fujikawa K., Galbany-Casals M., Garcia-Jacas
1349 N., Im H.-T., Kim S.-C., Liu J.-Q., López-Alvarado J., López-Pujol J., Mandel J.R.,
1350 Massó S., Mehregan I., Montes-Moreno N., Pyak E., Roquet C., Sáez L., Sennikov A.,
1351 Susanna A., Vilatersana R. 2018. Exploring data processing strategies in NGS target
1352 enrichment to disentangle radiations in the tribe Cardueae (Compositae). *Mol.*
1353 *Phylogenet. Evol.* 128:69–87.
- 1354 Hoffman A.D. 1932. The Douglas Canyon flora of east central Washington. *J. Geol.* 40:735–738.
- 1355 Hohenlohe P.A., Amish S.J., Catchen J.M., Allendorf F.W., Luikart G. 2011. Next-generation
1356 RAD sequencing identifies thousands of SNPs for assessing hybridization between
1357 rainbow and westslope cutthroat trout: SNP Discovery: Next Generation Sequencing.
1358 *Mol. Ecol. Resour.* 11:117–122.

ZHOU ET AL.

- 1359 Hollick C.A. 1936. The tertiary floras of Alaska. Washington: U.S. Govt. Print. Off. p. 1–171.
- 1360 Hou Y., Nowak M.D., Mirré V., Bjorå C.S., Brochmann C., Popp M. 2016. RAD-seq data point
1361 to a northern origin of the arctic–alpine genus *Cassiope* (Ericaceae). *Mol. Phylogenet.*
1362 *Evol.* 95:152–160.
- 1363 Huzioka K. 1972. The tertiary floras of Korea. *J. Ming. Coll. Akita Univ.* 5(1): 1–83.
- 1364 Ishida S. 1970. The Noroshi flora of Note Peninsula, Central Japan. *Memoirs of the Faculty of*
1365 *Science, Kyoto University. Series of Geology and Mineralogy.* 37(1):1–112.
- 1366 Jansen R.K., Saski C., Lee S.-B., Hansen A.K., Daniell H. 2011. Complete Plastid Genome
1367 Sequences of Three Rosids (*Castanea*, *Prunus*, *Theobroma*): Evidence for At Least Two
1368 Independent Transfers of rpl22 to the Nucleus. *Mol. Biol. Evol.* 28:835–847.
- 1369 Jisaburō O., Meyer F.G., Walker E.H. 1965. *Flora of Japan*. Washington D.C.: Smithsonian
1370 Institution. p. 1–1067.
- 1371 Johnson G.P. 1988. Revision of *Castanea* sect *Balanocastanon* (Fagaceae). *J. Arnold Arbor.*:25–
1372 49.
- 1373 Johnson M.G., Gardner E.M., Liu Y., Medina R., Goffinet B., Shaw A.J., Zerega N.J.C., Wickett
1374 N.J. 2016. HybPiper: Extracting Coding Sequence and Introns for Phylogenetics from
1375 High-Throughput Sequencing Reads Using Target Enrichment. *Appl. Plant Sci.*
1376 4:1600016.
- 1377 Johnson M.G., Pokorny L., Dodsworth S., Botigué L.R., Cowan R.S., Devault A., Eiserhardt
1378 W.L., Epiawalage N., Forest F., Kim J.T., Leebens-Mack J.H., Leitch I.J., Maurin O.,
1379 Soltis D.E., Soltis P.S., Wong G.K., Baker W.J., Wickett N.J. 2019. A Universal Probe Set
1380 for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed
1381 Using k-Medoids Clustering. *Syst. Biol.* 68:594–606.
- 1382 Kates H.R., Johnson M.G., Gardner E.M., Zerega N.J.C., Wickett N.J. 2018. Allele phasing has
1383 minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in a
1384 case study of *Artocarpus*. *Am. J. Bot.* 105:404–416.
- 1385 Lakhanpal R.N. 1958. *The Rujada Flora of West Central Oregon*. University of California
1386 Publications in Geological Sciences. 35:1–65.
- 1387 Lambert S.M., Streicher J.W., Fisher-Reid M.C., Méndez de la Cruz F.R., Martínez-Méndez N.,
1388 García-Vázquez U.O., Nieto Montes de Oca A., Wiens J.J. 2019. Inferring introgression
1389 using RADseq and *D*_{FOIL}: Power and pitfalls revealed in a case study of spiny lizards
1390 (*Sceloporus*). *Mol. Ecol. Resour.* 19:818–837.
- 1391 Landis J.B., Soltis D.E., Li Z., Marx H.E., Barker M.S., Tank D.C., Soltis P.S. 2018. Impact of
1392 whole-genome duplication events on diversification rates in angiosperms. *Am. J. Bot.*
1393 105:348–363.

A NEW PARALOG REMOVAL PIPELINE

- 1394 Landis M.J., Matzke N.J., Moore B.R., Huelsenbeck J.P. 2013. Bayesian Analysis of
1395 Biogeography when the Number of Areas is Large. *Syst. Biol.* 62:789–804.
- 1396 Lang P., Dane F., Kubisiak T.L. 2006. Phylogeny of *Castanea* (Fagaceae) based on chloroplast
1397 trnT-L-F sequence data. *Tree Genet. Genomes.* 2:132–139.
- 1398 Lang P., Dane F., Kubisiak T.L., Huang H. 2007. Molecular evidence for an Asian origin and a
1399 unique westward migration of species in the genus *Castanea* via Europe to North
1400 America. *Mol. Phylogenet. Evol.* 43:49–59.
- 1401 Larridon I., Villaverde T., Zuntini A.R., Pokorny L., Brewer G.E., Eritawallage N., Fairlie I.,
1402 Hahn M., Kim J., Maguilla E., Maurin O., Xanthos M., Hipp A.L., Forest F., Baker W.J.
1403 2020. Tackling Rapid Radiations With Targeted Sequencing. *Front. Plant Sci.* 10:1655.
- 1404 Leache A.D., Chavez A.S., Jones L.N., Grummer J.A., Gottscho A.D., Linkem C.W. 2015.
1405 Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture
1406 versus restriction site associated DNA sequencing. *Genome Biol. Evol.* 7:706–719.
- 1407 Leaché A.D., Oaks J.R. 2017. The Utility of Single Nucleotide Polymorphism (SNP) Data in
1408 Phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* 48:69–84.
- 1409 Lecaudey L.A., Schlieven U.K., Osinov A.G., Taylor E.B., Bernatchez L., Weiss S.J. 2018.
1410 Inferring phylogenetic structure, hybridization and divergence times within Salmoninae
1411 (Teleostei: Salmonidae) using RAD-sequencing. *Mol. Phylogenet. Evol.* 124:82–99.
- 1412 Lemmon E.M., Lemmon A.R. 2013. High-Throughput Genomic Data in Systematics and
1413 Phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* 44:99–121.
- 1414 Leonard S. 2006. A New Species of Witch-Hazel (*Hamamelis*: Hamamelidaceae) Apparently
1415 Endemic to Southern Mississippi. *SIDA, Contributions to Botany.* 22(2):849–856.
- 1416 Lveill-Bourret ., Starr J.R., Ford B.A., Moriarty Lemmon E., Lemmon A.R. 2018. Resolving
1417 Rapid Radiations within Angiosperm Families Using Anchored Phylogenomics. *Syst.*
1418 *Biol.* 67:94–112.
- 1419 Lewin H.A., Robinson G.E., Kress W.J., Baker W.J., Coddington J., Crandall K.A., Durbin R.,
1420 Edwards S.V., Forest F., Gilbert M.T.P., Goldstein M.M., Grigoriev I.V., Hackett K.J.,
1421 Haussler D., Jarvis E.D., Johnson W.E., Patrinos A., Richards S., Castilla-Rubio J.C., van
1422 Sluys M.-A., Soltis P.S., Xu X., Yang H., Zhang G. 2018. Earth BioGenome Project:
1423 Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U.S.A.* 115:4325–4333.
- 1424 Lewis P.O. 2001. A likelihood approach to estimating phylogeny from discrete morphological
1425 character data. *Syst. Biol.* 50:913–925.
- 1426 Li H., Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.
1427 *Bioinformatics.* 25:1754–1760.

- 1428 Li J., Bogle A.L., Klein A.S., Donoghue M.J. 2000. PHYLOGENY AND BIOGEOGRAPHY OF
1429 HAMAMELIS (HAMAMELIDACEAE). Harv. Pap. Bot. 5:171–178.
- 1430 Li J., Bogle A.L. 2001. A New Suprageneric Classification System of the Hamamelidoideae
1431 Based on Morphology and Sequences of Nuclear and Chloroplast DNA. Harv. Pap. Bot.
1432 5(2):499–515.
- 1433 Limborg M.T., Seeb L.W., Seeb J.E. 2016. Sorting duplicated loci disentangles complexities of
1434 polyploid genomes masked by genotyping by sequencing. Mol. Ecol. 25:2117–2129.
- 1435 Lindenbaum P. (2015). msa2vcf utility within Jvarkit. Available:
1436 <https://github.com/lindenb/jvarkit>.
- 1437 Lischer H.E.L., Excoffier L., Heckel G. 2014. Ignoring Heterozygous Sites Biases Phylogenomic
1438 Estimates of Divergence Times: Implications for the Evolutionary History of Microtus
1439 Voles. Mol. Biol. Evol. 31:817–831.
- 1440 Long C., Kubatko L. 2018. The Effect of Gene Flow on Coalescent-based Species-Tree
1441 Inference. Syst. Biol. 67:770–785.
- 1442 Lu L.-M., Mao L.-F., Yang T., Ye J.-F., Liu B., Li H.-L., Sun M., Miller J.T., Mathews S., Hu H.-
1443 H., Niu Y.-T., Peng D.-X., Chen Y.-H., Smith S.A., Chen M., Xiang K.-L., Le C.-T., Dang
1444 V.-C., Lu A.-M., Soltis P.S., Soltis D.E., Li J.-H., Chen Z.-D. 2018. Evolutionary history
1445 of the angiosperm flora of China. Nature. 554:234–238.
- 1446 Luo A., Duchêne D.A., Zhang C., Zhu C.-D., Ho S.Y. 2020. A Simulation-Based Evaluation of
1447 Tip-Dating Under the Fossilized Birth–Death Process. Syst. Biol. 69:325–344.
- 1448 Ma Q., Zhang W., Jenny Xiang Q.-Y. 2017. Evolution and developmental genetics of floral
1449 display-A review of progress: A review of progress in evo-devo of floral display. J. Syst.
1450 Evol. 55:487–515.
- 1451 MacGinitie H.D. 1941. Middle Eocene flora from the central Sierra Nevada. Washington, D. C.:
1452 Carnegie Institution of Washington. p. 1–167.
- 1453 MacGinitie H.D. 1953. Fossil plants of the Florissant Beds, Colorado. Washington, D. C.:
1454 Carnegie Institution of Washington. p. 1–198.
- 1455 MacGuigan D.J., Near T.J. 2019. Phylogenomic Signatures of Ancient Introgression in a Rogue
1456 Lineage of Darters (Teleostei: Percidae). Syst. Biol. 68:329–346.
- 1457 Madlung A. 2013. Polyploidy and its effect on evolutionary success: old questions revisited with
1458 new tools. Heredity. 110:99–104.
- 1459 Malinsky M., Trucchi E., Lawson D.J., Falush D. 2018. RADpainter and fineRADstructure:
1460 Population Inference from RADseq Data. Mol. Biol. and Evol. 35:1284–1290.

A NEW PARALOG REMOVAL PIPELINE

- 1461 Malinsky M. 2019. Dsuite-fast D-statistics and related admixture evidence from VCF files.
1462 BioRxiv. DOI: <https://doi.org/10.1101/634477>.
- 1463 Mai D.H. 2001. Die mittelmiozänen und obermiozänen Floren aus der Meuroer und Raunoer
1464 Folge in der Lausitz. Teil III: Fundstellen und Palaeobiologie. Palaeontogr. Abt. B.
1465 258:1–85.
- 1466 Manchester S.R., Chen Z., Geng B., Tao J. 2005. Middle Eocene flora of Huadian, Jilin
1467 Province, Northeastern China. Acta Palaeobot. 45(1):3–26.
- 1468 Manchester S.R., Chen Z.-D., Lu A.-M., Uemura K. 2009. Eastern Asian endemic seed plant
1469 genera and their paleogeographic history throughout the Northern Hemisphere. J. Syst.
1470 Evol. 47:1–42.
- 1471 Mann P. 2007. Overview of the tectonic history of northern Central America. Geol. Soc. Am.
1472 Spec. Pap. 428:1–19.
- 1473 Manos P.S., Meireles J.E. 2015. Biogeographic analysis of the woody plants of the Southern
1474 Appalachians: Implications for the origins of a regional flora. Am. J. Bot. 102:780–804.
- 1475 Manthey J.D., Campillo L.C., Burns K.J., Moyle R.G. 2016. Comparison of Target-Capture and
1476 Restriction-Site Associated DNA Sequencing for Phylogenomics: A Test in Cardinalid
1477 Tanagers (Aves, Genus: *Piranga*). Syst. Biol. 65:640–650.
- 1478 Matzke N.J. 2013. Probabilistic historical biogeography: new models for founder-event
1479 speciation, imperfect detection, and fossils allow improved accuracy and model-testing.
1480 Front. Biogeogr. 5.4.
- 1481 McCartney D.L., Walker R.M., Morris S.W., McIntosh A.M., Porteous D.J., Evans K.L. 2016.
1482 Identification of polymorphic and off-target probe binding sites on the Illumina Infinium
1483 MethylationEPIC BeadChip. Genom. Data. 9:22–24.
- 1484 McCormack J.E., Hird S.M., Zellmer A.J., Carstens B.C., Brumfield R.T. 2013. Applications of
1485 next-generation sequencing to phylogeography and phylogenetics. Mol. Phylogenet.
1486 Evol. 66:526–538.
- 1487 McCormack J.E., Tsai W.L., Faircloth B.C. 2016. Sequence capture of ultraconserved elements
1488 from bird museum specimens. Mol. Ecol. Resour. 16:1189–1203.
- 1489 McKinney G.J., Waples R.K., Seeb L.W., Seeb J.E. 2017. Paralogs are revealed by proportion of
1490 heterozygotes and deviations in read ratios in genotyping-by-sequencing data from
1491 natural populations. Mol Ecol Resour. 17:656–669.
- 1492 Medina R., Johnson M.G., Liu Y., Wickett N.J., Shaw A.J., Goffinet B. 2019. Phylogenomic
1493 delineation of Physcomitrium (Bryophyta: Funariaceae) based on targeted sequencing of
1494 nuclear exons and their flanking regions rejects the retention of Physcomitrella,
1495 Physcomitridium and Aphanorrhagma. J. Syst. Evol. 57:404–417.

- 1496 Miller M. A., Pfeiffer W., Schwartz T., 2010, Creating the CIPRES Science Gateway for
1497 inference of large phylogenetic trees. In *Gateway Computing Environments Workshop*
1498 (*GCE*). IEEE. p. 1–8. DOI: 10.1109/GCE.2010.5676129.
- 1499 Milne R.L. 2006. Northern Hemisphere Plant Disjunctions: A Window on Tertiary Land Bridges
1500 and Climate Change? *Ann. Bot.* 98:465–472.
- 1501 Nixon K.C., Crepet W.L. 1989. *Trigonobalanus* (Fagaceae): taxonomic status and phylogenetic
1502 relationships. *Am. J. Bot.* 76:828–841.
- 1503 Mu X.-Y., Tong L., Sun M., Zhu Y.-X., Wen J., Lin Q.-W., Liu B. 2020. Phylogeny and
1504 divergence time estimation of the walnut family (Juglandaceae) based on nuclear RAD-
1505 Seq and chloroplast genome data. *Mol. Phylogenet. Evol.* 147:106802.
- 1506 Murphy B., Forest F., Barraclough T., Rosindell J., Bellot S., Cowan R., Golos M., Jebb M.,
1507 Cheek M. 2020. A phylogenomic analysis of *Nepenthes* (Nepenthaceae). *Mol.*
1508 *Phylogenet. Evol.* 144:106668.
- 1509 Nguyen L.-T., Schmidt H.A., Von Haeseler A., Minh B.Q. 2015. IQ-TREE: a fast and effective
1510 stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*
1511 32:268–274.
- 1512 Nixon K. 1997. *Castanea*. In *Flora of North America North of Mexico*. New York: Oxford
1513 University Press. 3:439–442.
- 1514 Ohwi J. 1978. *Hamamelis*. In: *Flora of Japan*. Tokyo: Shibundo Co. Ltd. Publishers. p. 1–724.
- 1515 One Thousand Plant Transcriptomes Initiative. 2019. One thousand plant transcriptomes and the
1516 phylogenomics of green plants. *Nature*. 574:679–685.
- 1517 Ozaki K. 1991. Late Miocene and Pliocene floras in central Honshu, Japan. Yokohama, Japan:
1518 Kanagawa Prefectural Museum. p. 1–244.
- 1519 Pais A.L., Li X., Jenny Xiang Q.-Y. 2018. Discovering variation of secondary metabolite
1520 diversity and its relationship with disease resistance in *Cornus florida* L. *Ecol. Evol.*
1521 8:5619–5636.
- 1522 Pais A.L., Whetten R.W., Xiang Q. (Jenny). 2020. Population structure, Landscape Genomics,
1523 and Genetic Signatures of Adaptation to Exotic Disease Pressure in *Cornus florida* L.—
1524 Insights from GWAS and GBS data. *J. Syst. Evol.* 58(5): 546–570.
- 1525 Pais A.L., Whetten R.W., Xiang Q.Y.J. 2017. Ecological genomics of local adaptation in *Cornus*
1526 *florida* L. by genotyping by sequencing. *Ecol. Evol.* 7:441–465.
- 1527 Peterson B.K., Weber J.N., Kay E.H., Fisher H.S., Hoekstra H.E. 2012. Double digest RADseq:
1528 An inexpensive method for de novo SNP discovery and genotyping in model and non-
1529 model species. *PLoS ONE*. 7:e37135–e37135.

A NEW PARALOG REMOVAL PIPELINE

- 1530 Quinlan A.R., Hall I.M. 2010. BEDTools: a flexible suite of utilities for comparing genomic
1531 features. *Bioinformatics*. 26:841–842.
- 1532 Radtke M.G., Pigg K.B., Wehr W.C. 2005. Fossil *Corylopsis* and *Fothergilla* leaves
1533 (Hamamelidaceae) from the lower Eocene flora of Republic, Washington, USA, and their
1534 evolutionary and biogeographic significance. *Int. J. Plant Sci.* 166:347–356.
- 1535 Rambaut A. 2012. FigTree v1. 4. Available:
1536 <https://github.com/rambaut/figtree/releases/tag/v1.4.4>.
- 1537 Rambaut A., Drummond A.J., Xie D., Baele G., Suchard M.A. 2018. Posterior summarization in
1538 Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* 67:901–904.
- 1539 Reaz R., Bayzid M.S., Rahman M.S. 2014. Accurate phylogenetic tree reconstruction from
1540 quartets: a heuristic approach. *PLoS One*. 9:e104008.
- 1541 Ree R.H., Sanmartín I. 2018. Conceptual and statistical problems with the DEC+J model of
1542 founder-event speciation and its comparison with DEC via model selection. *J. Biogeogr.*
1543 45:741–749.
- 1544 Ree R.H., Smith S.A. 2008. Maximum likelihood inference of geographic range evolution by
1545 dispersal, local extinction, and cladogenesis. *Syst. Biol.* 57:4–14.
- 1546 Rognes T., Flouri T., Nichols B., Quince C., Mahé F. 2016. VSEARCH: a versatile open source
1547 tool for metagenomics. *PeerJ*. 4:e2584.
- 1548 Ronquist F., Cannatella D. 1997. Dispersal-Vicariance Analysis: A New Approach to the
1549 Quantification of Historical Biogeography. *Syst. Biol.* 46:195–203.
- 1550 Ronquist F., Huelsenbeck J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed
1551 models. *Bioinformatics*. 19:1572–1574.
- 1552 Ronquist F., Klopfstein S., Vilhelmsen L., Schulmeister S., Murray D.L., Rasnitsyn A.P. 2012. A
1553 total-evidence approach to dating with fossils, applied to the early radiation of the
1554 Hymenoptera. *Syst. Biol.* 61:973–999.
- 1555 Sakaguchi S., Takahashi D., Setoguchi H., Isagi Y. 2018. Genetic structure of the clonal herb
1556 *Tanakaea radicans* (Saxifragaceae) at multiple spatial scales, revealed by nuclear and
1557 mitochondrial microsatellite markers. *Plant Species Biol.* 33:81–87.
- 1558 Sargent C.S. 1890. Hamamelidaceae-Sapotaceae. In *The sylvia of North America*. New York:
1559 Peter Smith.
- 1560 Shee Z.Q., Frodin D.G., Cámara-Leret R., Pokorny L. 2020. Reconstructing the Complex
1561 Evolutionary History of the Papuan Schefflera Radiation Through Herbariomics.
1562 *Front. Plant Sci.* 11:258.

ZHOU ET AL.

- 1563 Simpson J.T., Wong K., Jackman S.D., Schein J.E., Jones S.J.M., Birol I. 2009. ABySS: A
1564 parallel assembler for short read sequence data. *Genome Res.* 19:1117–1123.
- 1565 Sims H.J., Herendeen P.S., Crane P.R. 1998. New genus of fossil Fagaceae from the Santonian
1566 (Late Cretaceous) of central Georgia, USA. *Int. J. Plant Sci.* 159:391–404.
- 1567 Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of
1568 large phylogenies. *Bioinformatics.* 30:1312–1313.
- 1569 Standley P.C. 1937. Studies of American plants, VII. *Field Mus. Nat. Hist., Bot. ser.* 17:155–224.
- 1570 Swofford D.L. 2003. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods)
1571 Version 4. Sinauer Associates, Sunderland, Massachusetts. Available:
1572 <https://paup.phylosolutions.com/>.
- 1573 Tanai T. 1970. The Oligocene floras from the Kushiro coal field, Hokkaido, Japan. *J. Fac. Sci.,*
1574 *Hokkaido Univ., Ser. 4, Geology and mineralogy.* 14:383–514.
- 1575 Tanai T. 1995. Fagacean Leaves from the Paleogene of Hokkaido, Japan. *Bull. Natl. Mus. Nat.*
1576 *Sci. Ser. C.* 21:71–102.
- 1577 Than C., Ruths D., Nakhleh L. 2008. PhyloNet: a software package for analyzing and
1578 reconstructing reticulate evolutionary relationships. *BMC Bioinform.* 9:322.
- 1579 Tiffney B.H. 1985a. Perspectives on the origin of the floristic similarity between eastern Asia
1580 and eastern North America. *J. Arnold Arbor.* 66:73–94.
- 1581 Tiffney B.H. 1985b. The Eocene North Atlantic land bridge: its importance in Tertiary and
1582 modern phytogeography of the Northern Hemisphere. *J. Arnold Arbor.* 66:243–273.
- 1583 Tiffney B.H., Manchester S.R. 2001. The Use of Geological and Paleontological Evidence in
1584 Evaluating Plant Phylogeographic Hypotheses in the Northern Hemisphere Tertiary. *Int.*
1585 *J. Plant Sci.* 162:S3–S17.
- 1586 Van der Burgh J. 1983. Allochthonous seed and fruit floras from the Pliocene of the Lower Rhine
1587 Basin. *Rev. Palaeobot. Palynol.* 40:33–90.
- 1588 Weitemier K., Straub S.C.K., Cronn R.C., Fishbein M., Schmickl R., McDonnell A., Liston A.
1589 2014. Hyb-Seq: Combining Target Enrichment and Genome Skimming for Plant
1590 Phylogenomics. *Appl. Plant Sci.* 2:1400042.
- 1591 Wen J. 1999. Evolution of Eastern Asian and Eastern North American Disjunct Distributions in
1592 Flowering Plants. *Annu. Rev. Ecol. Syst.* 30:421–455.
- 1593 Wen J., Ickert-Bond S., Nie Z.-L., Li R. 2010. Timing and modes of evolution of eastern Asian-
1594 North American biogeographic disjunctions in seed plants. In *Darwin's heritage today:*
1595 *Proceedings of the Darwin 2010 Beijing international conference.* Beijing: Higher
1596 Education Press. p. 252–269.

A NEW PARALOG REMOVAL PIPELINE

- 1597 Wen J., Nie Z., Ickert-Bond S.M. 2016. Intercontinental disjunctions between eastern Asia and
1598 western North America in vascular plants highlight the biogeographic importance of the
1599 Bering land bridge from late Cretaceous to Neogene. *J. Syst. Evol.* 54:469–490.
- 1600 Wen J., Shi S. 1999. A phylogenetic and biogeographic study of *Hamamelis* (Hamamelidaceae),
1601 an eastern Asian and eastern North American disjunct genus. *Biochem. Syst. Ecol.* 27:55–
1602 66.
- 1603 Wetterstrand K.A. 2015. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing
1604 Program (GSP). Available at: www.genome.gov/sequencingcosts.
- 1605 Wolfe J.A. 1980. Tertiary climates and floristic relationships at high latitudes in the Northern
1606 Hemisphere. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 30:313–323.
- 1607 Xiang Q. (Jenny), Soltis D.E. 2001. Dispersal-Vicariance Analyses of Intercontinental Disjuncts:
1608 Historical Biogeographical Implications for Angiosperms in the Northern Hemisphere.
1609 *Int. J. Plant Sci.* 162:S29–S39.
- 1610 Xiang Q.-Y., Crawford D.J., Wolfe A.D., Tang Y.-C., DePamphilis C.W. 1998a. Origin and
1611 Biogeography of *Aesculus* L. (Hippocastanaceae): A Molecular Phylogenetic Perspective.
1612 *Evolution.* 52.
- 1613 Xiang Q.-Y., Soltis D.E., Soltis P.S. 1998b. The Eastern Asian and Eastern and Western North
1614 American Floristic Disjunction: Congruent Phylogenetic Patterns in Seven Diverse
1615 Genera. *Mol. Phylogenet. Evol.* 10:178–190.
- 1616 Xiang Q.-Y., Soltis D.E., Soltis P.S., Manchester S.R., Crawford D.J. 2000. Timing the Eastern
1617 Asian–Eastern North American Floristic Disjunction: Molecular Clock Corroborates
1618 Paleontological Estimates. *Mol. Phylogenet. Evol.* 15:462–472.
- 1619 Xie L., Yi T.-S., Li R., Li D.-Z., Wen J. 2010. Evolution and biogeographic diversification of the
1620 witch-hazel genus (*Hamamelis* L., Hamamelidaceae) in the Northern Hemisphere. *Mol.*
1621 *Phylogenet. Evol.* 56:675–689.
- 1622 Yu G., Lam T.T.-Y., Zhu H., Guan Y. 2018. Two Methods for Mapping and Visualizing
1623 Associated Data on Phylogeny Using *Ggtree*. *Mol. Biol. Evol.* 35:3041–3043.
- 1624 Yu Y., Harris A.J., Blair C., He X. 2015. RASP (Reconstruct Ancestral State in Phylogenies): A
1625 tool for historical biogeography. *Mol. Phylogenet. Evol.* 87:46–49.
- 1626 Yu Y., Nakhleh L. 2015. A maximum pseudo-likelihood approach for phylogenetic networks.
1627 *BMC Genom.* 16:S10.
- 1628 Zarza E., Faircloth B.C., Tsai W.L.E., Bryson R.W., Klicka J., McCormack J.E. 2016. Hidden
1629 histories of gene flow in highland birds revealed with genomic markers. *Mol. Ecol.*
1630 25:5144–5157.

ZHOU ET AL.

- 1631 Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: polynomial time species tree
1632 reconstruction from partially resolved gene trees. BMC bioinform. 19:153.
- 1633 Zhang Z., Lu A. 1995. Hamamelidaceae: geographic distribution, fossil history and origin.
1634 Chinese Science Abstracts Series B. 6:37.
- 1635 Zhou W., Ji X., Obata S., Pais A., Dong Y., Peet R., Xiang Q.J. 2018. Resolving relationships and
1636 phylogeographic history of the *Nyssa sylvatica* complex using data from RAD-seq and
1637 species distribution modeling. Mol. Phylogenet. Evol. 126:1–16.
- 1638 Zhou W., Xiang Q.-Y., Wen J. 2020. Phylogenomics, biogeography, and evolution of
1639 morphology and ecological niche of the eastern Asian–eastern North American *Nyssa*
1640 (*Nyssaceae*). J. of Syst. Evol. 58:571–603.
- 1641 Zimmer E.A., Wen J. 2015. Using nuclear gene data for plant phylogenetics: Progress and
1642 prospects II. Next-gen approaches: Nuclear data for plant phylogenetics II. J. of Syst.
1643 Evol. 53:371–379.
- 1644

A NEW PARALOG REMOVAL PIPELINE

Figure Captions:

Fig. 1. Concepts of ortholog-paralog and the key steps of workflow of current bioinformatics programs 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 (Johnson et al. 2016), including reads mapping using BLASTx or BWA (Li and Durbin 2009), de novo assembly of contigs using SPAdes (Bankevich et al. 2012) and paralogs (in red) detection. c) The workflow of PHYLUCE (Faircloth 2016) and SECAPR (Andermann et al. 2018a), including de novo assembly of contigs from reads using SPAdes (in PHYLUCE) and Abyss (Simpson et al. 2009) (in SECAPR), respectively, contigs mapping using LASTZ (Harris 2007), and detection of paralogs (in red), followed by removal of detected paralogs. All three methods detect putative paralogs by observation of more than one unique contigs in an individual mapped to the same reference gene with >85% (HybPiper) or >80% (PHYLUCE and SECAPR) sequence identity or the same contig mapped to more than one reference genes.

Fig. 2. The putative paralogs detection (PPD) pipeline analytical workflow. The flowchart shows the basic PPD functions, including two major steps: first, generating a “degenerated” sequences; second, generating well-trimmed matrix and detecting paralogous genes. The “well-trimmed” process (s1 to s7) include filtering out sequences with high percentage of heterozygous sites (>5%), constructing Multiple Sequence Alignment (MSA), trimming MSA by removing sites missing in >50% samples (gaps) and hypervariable sites by a sliding window trimming method. The paralog detection step (s8) can detect additional paralogs indicated by presence of heterozygous site(s) present in >50% samples in the MSA. The “*degenerated*” sequence means

the sequence were coded by IUPAC ambiguity codes. See details in Supplementary Information, available on Dryad. Sp: species; OG: Outgroup.

Fig. 3. Gene trees and species tree of *Castanea* derived from different data sets. a) (1): the concatenated gene tree from RAD-seq M50 dataset; (2): the concatenated gene tree from well-trimmed “BWA100 *degenerated*” supercontig of orthologous gene data from Angiosperm 353 Hyb-Seq (primary Hyb-Seq-PPD data). Dash lines in the middle connecting the same samples used in both RAD-seq and Hyb-Seq analyses. Both concatenated gene trees in A were derived from analyses using the IQ-TREE whose topology is identical to those from RAxML and IQ-Tree. Numbers on branches are values of UF bootstrap from the IQ-TREE (only >90% numbers were shown), bootstrap values from RAxML (only >50% numbers were shown) (Supplementary Fig. S8, available on Dryad), and posterior probabilities from MrBayes (Supplementary Fig. S9, available on Dryad), respectively. Asterisks represent 100% support values. b) ML trees from 12,647 bp of plastid gene captured in the Hyb-Seq data by RAxML and IQ-Tree. Numbers on branches are combined from bootstrap in RAxML and ultrafast bootstrap in IQ-Tree. c) The combined species tree from RAD-seq and Hyb-Seq data derived from analyses using ASTRAL-III and SVDQuartets. The numbers on the branches are support values from the RAD-seq data with ASTRAL-III, RAD-seq data with SVDQuartets, Hyb-Seq data with ASTRAL-III, and Hyb-Seq data with SVDQuartets, respectively. *Quercus castaneifolia* was used as the outgroup in all analysis. All trees were drawn by ggtree in R.

Fig. 4. Gene trees and species tree of *Hamamelis* derived from different data sets. a) (1): the concatenated gene tree from RAD-seq M50 data; (2): the concatenated gene tree from the

A NEW PARALOG REMOVAL PIPELINE

primary Hyb-Seq-PPD data. Dash lines in the middle connecting the same samples used in both RAD-seq and Hyb-Seq approaches. Both concatenated gene trees in A were derived from analyses using the IQ-TREE, whose topology is identical to those from RAxML and IQ-Tree. Numbers on branches are values of UF bootstrap from IQ-Tree (only >90% numbers were shown), bootstrap values from RAxML (only >50% numbers were shown) (Supplementary Fig. S18, available on Dryad), and posterior probabilities from MrBayes (Supplementary Fig. S19, available on Dryad), respectively. Asterisks represent 100% support values. b) ML tree from 107,068 bp of plastid gene from Hyb-Seq data derived from analyses using RAxML and IQ-Tree. Numbers on branches are bootstrap values from analyses using RAxML and IQ-Tree, respectively. c) The combined species tree from RAD-seq and Hyb-Seq data derived from analyses using ASTRAL and SVDQuartets. Numbers on the branches are support values from ASTRAL-III (RAD-seq), ASTRAL-III (Hyb-Seq), and SVDQuartets (Hyb-Seq), respectively. The species tree from RAD-seq data with SVDQuartets method shows a different species tree and is presented in Supplementary Fig. S20, available on Dryad. *Parrotiopsis jacquemontiana* and *Fothergilla gardenii* were used as outgroups. All trees were drawn by ggtree in R.

Fig. 5 Results of divergence times estimated using RAD-seq and Hyb-Seq data. a) and b), results from RAD-seq M50 data; c) and d), results from the primary Hyb-Seq-PPD data. e) and f), results from well-trimmed “consensus” Hyb-Seq supercontig data. All topology was drawn by ggtree in R.

Fig. 6. Best supported species network of the orthologous genes from primary Hyb-Seq-PPD data inferred with PhyloNet. Numbers next to the hybrid branches indicate inheritance probabilities.

Fig. 7 Results of biogeographic analyses of *Castanea* and *Hamamelis* using the DEC model based on dated phylogeny using total evidence with combined RAD-Hyb-Seq data and morphological data. a) DEC result of *Castanea*. b) DEC result of *Hamamelis*. Letters at each node indicate the most-likely ancestral distributions inferred from the analyses, whose probabilities are shown by non-black colors in the pie charts. The black color in the pie chart indicates the total proportion (likelihood) of all other alternative ancestral ranges in the results of DEC analyses. Inferred intercontinental dispersals are indicated by red arrows. Fossil taxa are labeled in gray color. c) Distribution and hypothesized migration routes of *Castanea* including living species and fossils. d) Distribution and hypothesized migration routes of *Hamamelis* including living species and fossils. A: eastern Asia; B: eastern North America and Mexico; C: western North America; E: Europe.

Fig. S1-Fig. S7 Concatenated gene trees of *Castanea* from RAD-seq datasets. These figures show the phylogenetic trees derived from analyses of M20 to M80 datasets, respectively, using IQ-Tree. *Quercus castaneifolia* was used as an outgroup. All topologies are identical, showing two monophyletic groups, EA clade and ENA clade. Numbers on branches indicate values of UF bootstrap support.

A NEW PARALOG REMOVAL PIPELINE

1734 Fig. S8 Concatenated gene tree of *Castanea* from analysis of a) RAD-seq M50 and b)
1735 primary Hyb-Seq-PPD data using RAxML. *Quercus castaneifolia* was used as an outgroup.
1736 Numbers on branches are values of bootstrap support.

1737
1738 Fig. S9 Concatenated gene tree of *Castanea* from analysis of a) RAD-seq M50 and b)
1739 primary Hyb-Seq-PPD data using MrBayes. *Quercus castaneifolia* was used as an outgroup.
1740 Numbers on branches are values of posterior probability.

1741
1742 Fig. S10 The species tree of *Castanea* from combined RAD-Hyb-Seq data derived from
1743 analyses using ASTRAL-III and SVDQuartets. The species tree result is congruent with the
1744 nuclear gene topology in Fig. 3. The numbers on the branches are support values from the
1745 ASTRAL-III and SVDQuartets, respectively. *Quercus castaneifolia* was used as an outgroup.

1746
1747 Fig. S11-Fig. S17 Concatenated gene trees of *Hamamelis* from analysis of RAD-seq data.
1748 These figures show the phylogenetic results from analyses of M20 to M80 datasets, respectively,
1749 using IQ-Tree. *Parrotiopsis jacquemontiana* and *Fothergilla major* were used as outgroups, but
1750 the trees were rerooted by the hybrid *Hamamelis x intermedia*. Numbers on branches are values
1751 of UF bootstrap support.

1752
1753 Fig. S18 Concatenated gene tree of *Hamamelis* from analysis of a) RAD-seq M50 and b)
1754 primary Hyb-Seq-PPD data using RAxML. *Parrotiopsis jacquemontiana* and *Fothergilla major*
1755 were used as outgroups. Numbers on branches are values of bootstrap support.

1756

Fig. S19 Concatenated gene tree of *Hamamelis* from analysis of a) RAD-seq M50 and b) primary Hyb-Seq-PPD data using MrBayes. *Parrotiopsis jacquemontiana* and *Fothergilla major* were used as outgroups. Numbers on branches are values of posterior probability.

Fig. S20 The species tree of *Hamamelis* from analysis of RAD-seq M50 data using SVDQuartets. *Parrotiopsis jacquemontiana* and *Fothergilla major* were used as outgroups. The result shows that EA clade is a monophyly with relatively low support value 62.

Fig. S21 Species trees of *Hamamelis* from combined RAD-Hyb-Seq data derived from analyses using ASTRAL-III and SVDQuartets. Two species trees were identical except the ENA clade. The numbers on the branches are support values from the ASTRAL-III and SVDQuartets, respectively. The SVDQuartets result of ENA clade was shown in the insert. *Parrotiopsis jacquemontiana* and *Fothergilla gardenii* were used as outgroups.

Fig. S22 Divergence times resulting from well-trimmed “BWA100 degenerated” supercontig of orthologous genes from combined RAD-Hyb-Seq data of a) *Castanea* and b) *Hamamelis*. All results showed a similar divergence time results with that from degenerated superontig result from the primary Hyb-Seq data in Fig. 5.

Fig. S23 Comparisons of phylogenies derived from “consensus” and “degenerated” matrices of the Hyb-Seq data. The first two rows show the results from *Castanea* exon and supercontig regions and the last two rows show the results from *Hamamelis* exon and supercontig regions. The columns show the results from untrimmed “consensus”, well-trimmed “consensus”,

A NEW PARALOG REMOVAL PIPELINE

and well-trimmed “*degenerated*”, respectively. The results indicated the tree topologies are identical in both “*consensus*” and “*degenerated*”, but “*consensus*” matrices have longer branch lengths than “*degenerated*” matrices. The ENA clades in both *Castanea* and *Hamamelis* are highlighted by blue color, EA clades are highlighted by read, and European clades are highlighted by yellow.

Fig. S24 Comparisons of phylogenies derived from *orthologs* vs. *paralogs*, exon vs. supercontig, matrices using different trimmed matrices and different mapping methods of the Hyb-Seq data. In orthologs and paralogs comparison, the result shows the paralogs in exon regions have different tree topologies with orthologs in both *Castanea* and *Hamamelis*. “well-trimmed” matrices have shorter branch length than “untrimmed” matrices (also shown in Table S5). “auto-trimmed” matrices have shortest branch lengths and different tree topologies in paralogs. The ENA clades in both *Castanea* and *Hamamelis* are highlighted by blue color, EA clades are highlighted by read, and European clades are highlighted by yellow.

Fig. S25 Divergence times resulting from “BWA100 degenerated” paralogous supercontig, “well-trimmed degenerated” exon, and “auto-trimmed degenerated” supercontig matrices in Hyb-Seq. All results showed a similar divergence time results with that from degenerated supercontig result from Hyb-Seq in Fig. 5. But the “BWA100 degenerated” paralogous supercontig results show much older divergence times at shallow nodes (species level) (also shown in Table S5).

Fig. S26 Divergence times resulting from untrimmed Hyb-Seq “*consensus*” supercontig of orthologous genes of *Castanea* and *Hamamelis*. All results showed much older divergence time

1803 results with that from well-trimmed “*degenerated*” and “*consensus*” superontig results from Hyb-
1804 Seq data in Fig. 5.

1805
1806 Fig. S27 Results of biogeographic analyses of *Castanea* and *Hamamelis* using the DEC
1807 model based on Hyb-Seq data. a). Most likely ancestral ranges reconstructed for *Castanea* based
1808 on the total-evidence phylogeny with *Quercus castaneifolia* as outgroup. b). Most likely ancestral
1809 ranges reconstructed for *Hamamelis* based on the total-evidence phylogeny with *Parrotiopsis*
1810 *jacquemontiana* and *Fothergilla gardenii* as outgroups. A: eastern Asia; B: eastern North America
1811 and Mexico; C: western North America; E: Europe. The letters at each node indicate the most-
1812 likely ancestral distributions inferred from the analyses, whose probabilities are shown by non-
1813 black colors in the pie charts. The black color in the pie chart indicates the total proportion
1814 (likelihood) of all other alternative ancestral ranges in the results of DEC analyses. Intercontinental
1815 dispersal events inferred were indicated by red arrows. Fossil taxa are labeled in gray color.