Annals of Botany **129**: 457–471, 2022
https://doi.org/10.1093/aob/mcac007, available online at www.academic.oup.com/aob

ANNALS OF BOTANY
Founded 1887

Phylotranscriptomics of Theaceae: generic-level relationships, reticulation and whole-genome duplication

Qiong Zhang[1,2], Lei Zhao[3], Ryan A. Folk[4], Jian-Li Zhao[5], Nelson A. Zamora[6], Shi-Xiong Yang[1], Douglas E. Soltis[7], Pamela S. Soltis[7], Lian-Ming Gao[1,8], Hua Peng[1,*] and Xiang-Qin Yu[1,*]

[1]CAS Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, China, [2]College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China, [3]Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan 650201, China, [4]Department of Biological Sciences, Mississippi State University, MS 39762, USA, [5]Yunnan Key Laboratory of Plant Reproductive Adaptation and Evolutionary Ecology, Yunnan University, Kunming 650091, China, [6]National Herbarium of Costa Rica (CR), Natural History Department of National Museum of Costa Rica, San José, Costa Rica, [7]Florida Museum of Natural History, University of Florida, Gainesville, FL 32611, USA and [8]Yunnan Lijiang Forest Ecosystem National Observation and Research Station, Kunming Institute of Botany, Chinese Academy of Sciences, Lijiang 674100, Yunnan, China
** For correspondence. E-mail yuxiangqin@mail.kib.ac.cn and hpeng@mail.kib.ac.cn*

Received: 7 October 2021 Returned for revision: 29 December 21 Editorial decision: 12 January 2022 Accepted: 16 January 2022
Electronically published: 17 January 2022

• **Background and Aims** Theaceae, with three tribes, nine genera and more than 200 species, are of great economic and ecological importance. Recent phylogenetic analyses based on plastomic data resolved the relationships among the three tribes and the intergeneric relationships within two of those tribes. However, generic-level relationships within the largest tribe, Theeae, were not fully resolved. The role of putative whole-genome duplication (WGD) events in the family and possible hybridization events among genera within Theeae also remain to be tested further.
• **Methods** Transcriptomes or low-depth whole-genome sequencing of 57 species of Theaceae, as well as additional plastome sequence data, were generated. Using a dataset of low-copy nuclear genes, we reconstructed phylogenetic relationships using concatenated, species tree and phylogenetic network approaches. We further conducted molecular dating analyses and inferred possible WGD events by examining the distribution of the number of synonymous substitutions per synonymous site (*Ks*) for paralogues in each species. For plastid protein-coding sequences , phylogenies were reconstructed for comparison with the results obtained from analysis of the nuclear dataset.
• **Results** Based on the 610 low-copy nuclear genes (858 606 bp in length) investigated, Stewartieae was resolved as sister to the other two tribes. Within Theeae, the *Apterosperma–Laplacea* clade grouped with *Pyrenaria*, leaving *Camellia* and *Polyspora* as sister. The estimated ages within Theaceae were largely consistent with previous studies based mainly on plastome data. Two reticulation events within *Camellia* and one between the common ancestor of *Gordonia* and *Schima* were found. All members of the tea family shared two WGD events, an older At-γ and a recent Ad-β; both events were also shared with the outgroups (Diapensiaceae, Pentaphylacaceae, Styracaceae and Symplocaceae).
• **Conclusions** Our analyses using low-copy nuclear genes improved understanding of phylogenetic relationships at the tribal and generic levels previously proposed based on plastome data, but the phylogenetic position of the *Apterosperma–Laplacea* clade needs more attention. There is no evidence for extensive intergeneric hybridization within Theeae or for a Theaceae-specific WGD event. Land bridges (e.g. the Bering land bridge) during the Late Oligocene may have permitted the intercontinental plant movements that facilitated the putative ancient introgression between the common ancestor of *Gordonia* and *Schima*.

Key words: Theaceae, phylogeny, transcriptome, low-copy nuclear genes, molecular dating, phylogenetic network, whole-genome duplication.

INTRODUCTION

Theaceae, the tea family, comprise nine genera in three tribes and contain 372 accepted species (WFO, 2021) of evergreen and deciduous trees and shrubs. Members of Theaceae have great economic and ecological importance; the family contains familiar plants such as tea [e.g. *Camellia sinensis* (L.) Kuntze], oil plants (e.g. *C. oleifera* Abel) and a number of woody ornamentals (e.g. *C. japonica* L., *C. reticulata* Lindl.). Some large tree representatives (e.g. *Schima*) and small tree lineages (e.g. *Camellia*, *Stewartia*) are dominant or common species of the subtropical evergreen broadleaved forests in East Asia (Tang, 2015). Due to excessive collection and habitat destruction, several species (mainly species of *Camellia*) have been listed as (critically) endangered, including *C. fangchengensis* S. Ye Liang & Y. C. Zhong, *C. hekouensis* C. J. Wang & G. S.

Fan and *C. piquetiana* (Pierre) Sealy (IUCN, 2020). In addition, many new species (Orel, 2006; Orel and Wilson, 2010*b*, 2012; Orel *et al.*, 2013; Orel and Curry, 2015, 2019; Lee and Yang, 2019; Liu *et al.*, 2019, 2020*a*, *b*; Yu *et al.*, 2021) and subgeneric taxa (Orel and Wilson, 2010*a*; Orel *et al.*, 2014) have been described and published for *Camellia* in the past decade.

Since the establishment of Theaceae (Mirbel, 1813), the systematic boundaries of genera in the family as defined by morphology have changed significantly (Bentham and Hooker, 1862; Melchior, 1925; Takhtajan, 1997), from two genera to as many as six tribes and 32 genera. Molecular systematic studies recognized three tribes and nine genera (Stevens, 2001 onwards; APG IV, 2016). Since then, a number of systematic studies using DNA markers, morphology, anatomy and cytology have been conducted to explore the relationships among tribes and genera (Ye, 1990; Tsou, 1998; Prince and Parks, 2001; S. X. Yang *et al.*, 2004; Wang *et al.*, 2006; J. B. Yang *et al.*, 2006; M. M. Li *et al.*, 2013; W. Zhang *et al.*, 2014; Yu *et al.*, 2017*b*). However, many phylogenetic relationships remain unresolved in this family.

First, phylogenetic relationships among the three tribes in Theaceae remain to be confirmed using nuclear genes. Evidence based on floral development indicated a sister relationship between tribes Gordonieae and Stewartieae (Tsou, 1998; treated as subtribes in this study). The same topology was recovered based on the small single-copy region (SSC) of the plastome, but with conflicting support among maximum parsimony (MP), maximum likelihood (ML) and Bayesian inference (BI) analyses (M. M. Li *et al.*, 2013). However, phylogenetic analysis of 46 morphological characters supported a sister relationship between Gordonieae (=Schimeae) and Theeae (Wang *et al.*, 2006), which was also supported by phylogenetics of plastid *rbcL*, *matK* and *trnL-F*, mitochondrial *matR* and nuclear ribosomal internal transcribed spacer (nrITS) (Prince and Parks, 2001; S. X. Yang *et al.*, 2004; J. B. Yang *et al.*, 2006) and whole plastome sequences (Yu *et al.*, 2017*b*). In a recent study of Ericales based on 25 loci from the plastid, nuclear and mitochondrial genomes, Theeae and Gordonieae were also recovered as sisters, with Stewartieae as their sister (Rose *et al.*, 2018). However, the sampling of nuclear genes in previous studies was very limited, and further analyses that incorporate more regions of the nuclear genome are needed.

Intergeneric relationships within Theeae have been challenging to resolve, especially the phylogenetic position of *Apterosperma* and *Laplacea*. Since the time of its original description (Chang, 1976), *Apterosperma* has been formally placed in tribe Gordonieae (=Schimeae) and considered close to *Schima* and *Franklinia* based on similar morphological characters (Ye, 1990; Chang and Ren, 1998; Tsou, 1998). A combined molecular phylogenetic analysis based on nrITS, plastid *trnL-F* and mitochondrial DNA (mtDNA) *matR* sequence data placed *Apterosperma* as sister to all other genera in Theeae (S. X. Yang *et al.*, 2004). However, based on analyses of 46 morphological characters, *Apterosperma* was placed as sister to other Theeae (Wang *et al.*, 2006). Using five genomic regions (plastid: *atpI-H*, *matK*, *psbA5′R-ALS-11F*, *rbcL*; nuclear: *LEAFY*) and 30 species representing four of the five genera within Theeae, W. Zhang *et al.* (2014) found that *Apterosperma* formed a sister relationship with *Polyspora* in the chloroplast DNA (cpDNA) tree, but was placed within a clade comprising *Tutcheria* (=*Pyrenaria*) and *Parapyrenaria* (=*Pyrenaria*) in the *LEAFY* tree. In that study, *Camellia* and *Pyrenaria* were not monophyletic, and inconsistent phylogenetic placement of some species between the nuclear and plastid trees was proposed to be the result of widespread hybridization among genera in Theeae.

In contrast to *Apterosperma*, very few studies have included *Laplacea*. Prince and Parks (2001) recovered *Laplacea* within a clade comprising *Camellia*, *Tutcheria* (=*Pyrenaria*) and *Glyptocarpa* (=*Camellia*). In Yu *et al.* (2017*b*), *Apterosperma* and *Laplacea* were sisters with strong support [ML bootstrap support (MLBS) = 96 %, BI posterior probability (PP) = 1.0]. However, the *Apterosperma–Laplacea* clade grouped either with *Camellia–Polyspora* or with *Pyrenaria* with moderate support based on different partitions of the plastome, grouped with *Polyspora* with weak support in the nrITS dataset, or even resolved as an early-branching clade from the combined plastome and nrITS dataset.

Species in the tea family are disjunctly distributed in temperate, subtropical and tropical areas of eastern to south-eastern Asia, and eastern North America to Central and South America, i.e. an Amphi-Pacific disjunction (Kobuski, 1949, 1950; Prince, 1993; Stevens, 2001 onwards; Ming and Bartholomew, 2007). Several studies have been conducted to investigate the biogeographical history of Theaceae, and some focused on the Eastern North American–East Asian disjunct genus *Stewartia* (Prince, 2002; H. Y. Lin *et al.*, 2019). Others attempted to uncover the spatial–temporal history of the whole family. Based on two calibration points, the biogeographical reconstruction of M. M. Li *et al.* (2013) indicated that Theaceae originated in the late Cretaceous [~86 million years ago (Mya)] and started to diversify in the early Eocene (~49 Mya); species interchange of Gordonieae and Stewartieae between North America and eastern Asia may have been facilitated by the Bering land bridge. Our previous study also suggested a late Cretaceous origin (~91.6 Mya) of the family, but in contrast to M. M. Li *et al.* (2013), the crown of Theaceae was inferred as late Palaeocene (~57.3 Mya) (Yu *et al.*, 2017*b*). Theaceae were proposed to be of Indo-Malaysian origin as part of a broader phylogenetic study of Ericales (Rose *et al.*, 2018), although with relatively sparse sampling. However, a more recent study suggested a broad mid- to high-latitude Northern Hemispheric (e.g. Eurasia, Nearctic) origin of Theaceae (Yan *et al.*, 2021). Biogeographical analyses based on a large number of nuclear genes and strong taxon sampling are needed to resolve this issue.

Ancient whole-genome duplication (WGD), an important evolutionary force in plants, has been reported in the common ancestor of extant seed plants, extant angiosperms and core eudicots (At-γ) (Jiao *et al.*, 2011; Vekemans *et al.*, 2012), as well as at many other critical nodes across the green plant tree of life (Leebens-Mack *et al.*, 2019). WGD events were also found in the early history of numerous families, such as Asteraceae, Brassicaceae, Fabaceae, Poaceae and Rosaceae (Barker *et al.*, 2008; Schranz *et al.*, 2012; C. H. Huang *et al.*, 2016; Xiang *et al.*, 2017; Qiao *et al.*, 2019). Based on the kiwifruit (*Actinidia chinensis*) genome, researchers proposed a WGD event called Ad-β, which was shared by *Actinidia* and *Camellia* (T. Shi *et al.*, 2010; S. X. Huang *et al.*, 2013). However, using genome collinearity and also the MAPS pipeline, Ad-β was recently

mapped to the core Ericales (Leebens-Mack *et al.*, 2019), and more specifically to the clade comprising core Ericales, primuloids, polemonioids and Lecythidaceae (C. Zhang *et al.*, 2020). In Theaceae, whole-genome sequence data were only available for *Camellia*, and results for the placement of the WGD event from different studies were not in agreement with each other (Xia *et al.*, 2017, 2020; Wei *et al.*, 2018); more data and broader sampling are needed.

Phylotranscriptomics has become an important approach for plant phylogenetics (Soltis *et al.*, 2013; Yang and Smith, 2013; Wickett *et al.*, 2014; C. H. Huang *et al.*, 2016; Yu *et al.*, 2018; C. Zhang *et al.*, 2020) and has been widely used to explore diverse evolutionary questions, including the origin and early diversification of land plants (Wickett *et al.*, 2014), deep-level (among eight clades) angiosperm phylogeny (Zeng *et al.*, 2014), phylogenetic relationships within eudicots, asterids and rosids, respectively (L. Zhao *et al.*, 2016; Zeng *et al.*, 2017; C. Zhang *et al.*, 2020), and intergeneric relationships for several species-rich orders or families (Caryophyllales, Cupressaceae, Asteraceae, Brassicaceae, Fabaceae, Rosaceae and Pinaceae) (C. H. Huang *et al.*, 2015, 2016; Y. Yang *et al.*, 2015; Xiang *et al.*, 2017; Ran *et al.*, 2018; Mao *et al.*, 2019; Y. Y. Zhao *et al.*, 2021). Additionally, the 1000 Plant Transcriptomes Project (1KP), which sequenced transcriptomes from 1124 species representing the diversity of green plants, provided resolution across the green tree of life (Matasci *et al.*, 2014; Leebens-Mack *et al.*, 2019). Furthermore, comparing evidence from the plastid and nuclear genomes allows the detection of cytoplasmic introgression and other forms of hybridization (Calvo *et al.*, 2013; Folk *et al.*, 2017; Guo *et al.*, 2018; Morales-Briones *et al.*, 2018; Stubbs *et al.*, 2020).

Here, transcriptomes of 55 species of Theaceae were sequenced, and low-depth whole-genome sequencing was conducted for another two species because no fresh tissue was available; both nuclear and plastid genes were extracted. Plastid genes were integrated with plastome data from our previous study (Yu *et al.*, 2017*a*, *b*). We aim to reconstruct the nuclear phylogenetic framework and temporal history of the tea family, focusing on the relationships among tribes and genera and the phylogenetic position of *Apterosperma* and *Laplacea*. Furthermore, we also test the intergeneric hybridization hypothesis proposed previously and investigate whether a previously detected WGD event in *Camellia* is shared by other genera of Theaceae.

MATERIALS AND METHODS

Taxon sampling, transcriptomics and low-depth whole-genome sequencing

We collected 58 samples representing 57 species from all three tribes and all nine genera of Theaceae (Table 1). Fresh and healthy leaves of 27 samples were collected in the field and then frozen immediately in liquid nitrogen. For 29 samples, we used leaf tissue collected in the field and stored in a −80 °C freezer. Additionally, silica-dried leaves of *Stewartia malacodendron* and *Laplacea fruticosa* were used for low-depth whole-genome sequencing (at least 30× coverage).

Total RNA was extracted from the 56 samples of flash frozen leaves using the Spectrum Plant Total RNA Extraction Kit (Sigma-Aldrich, Burlington, MA, USA). Total genomic DNA was isolated from silica-dried leaves of *Stewartia malacodendron* and *Laplacea fruticosa* using the modified CTAB method (Doyle and Doyle, 1987). RNA sequencing and low-depth whole-genome sequencing library construction, Illumina HiseqXten sequencing, raw data cleaning and quality control were performed at Novogene (China). Additionally, plastid genome data were obtained from our previous studies (Yu *et al.*, 2017*a*, *b*). Six species from Diapensiaceae, Pentaphylacaceae, Styracaceae and Symplocaceae were used as outgroups (Table 1). The transcriptome data of these six species were downloaded from GenBank.

Based on the results from previous phylogenetic studies of Theacaee, *Pyrenaria* and *Stewartia* were treated in the broad sense according to the treatment in the Flora of China (Ming and Bartholomew, 2007). In our previous study (Yu *et al.*, 2017*b*), we found that *Laplacea grandis* grouped with *Gordonia lasianthus* and should therefore be moved into *Gordonia*; thus, here we use the name *Gordonia brenesii* H. Keng (=*Laplacea grandis*) following Grayum and Madrigal (2011).

Sequence assembly and orthologue identification for nuclear genes

Quality control for all the raw sequencing reads was performed using Fastp v0.20.1 (S. F. Chen *et al.*, 2018), including removal of adapters, reads containing N and reads with low quality scores (percentage of base ≤Q20). Trinity v2.8.4 was used to conduct *de novo* assembly of cleaned Illumina RNA sequencing reads of each species (Grabherr *et al.*, 2011; Haas *et al.*, 2013). After transcripts had been filtered with Transrate v1.0.3 (Smith-Unna *et al.*, 2016) and clustered using Corset v1.07 (Davidson and Oshlack, 2014), TransDecoder v5.3.0 (Haas *et al.*, 2013) with blastp was utilized to choose open reading frames. CD-HIT v4.6 (Li and Godzik, 2006) was then used to remove redundant contigs with a threshold of 0.99. Orthology inference was conducted following Yang and Smith's (2014) pipeline, setting min_taxon as 48 in each orthologue group.

For the data obtained from low-depth whole-genome sequencing of two species (*Stewartia malacodendron* and *Laplacea fruticosa*), we first *de novo* assembled each genome using Platanus v1.2.4 with default parameters (Kajitani *et al.*, 2014). Then RepeatMasker v4.1.1 (Tarailo-Graovac and Chen, 2009) and RepeatModeler (http://www.repeatmasker.org/) were used to identify tandem repeats and transposable element (TEs). We carried out gene annotation using *de novo* gene prediction in AUGUSTUS (Haas *et al.*, 2008) and homologue prediction using Exonerate (Slater and Birney, 2005) and then generated an integrated gene set using EVidenceModeler (Haas *et al.*, 2008). Preliminary gene trees were reconstructed using RAxML v8.2.12 (Stamatakis, 2014).

To reduce potentially misidentified orthologues, we carefully examined the individual gene trees obtained for the 631 putative orthologues and found the three tribes within Theaceae were not monophyletic in 21 of these gene trees. Because all three tribes within Theaceae were consistently monophyletic in

TABLE 1. *List of taxa sampled in this study, with voucher, Illumina reads and GenBank accessions; species names that are underlined represent those species selected for PhyloNet analysis*

Taxon	Voucher specimen	Sources	No. of reads (trimmed)	SRA number	Plastid genome
Stewartieae					
Stewartia calcicola	YXQ090	Yunnan, China	76 314 960	SRR14596892	KY406783
Stewartia cordifolia	YXQ144	Guangxi, China	101 158 084	SRR14596891	KY406775
Stewartia crassifolia	YXQ171	Hunan, China	88 878 696	SRR14596880	KY406766
Stewartia malacodendron	FLAS 260361	Alabama, USA	1 798 360 616	SRR14596869	KY406773
Stewartia ovata	18847*A	The Arnold Arboretum	100 544 128	SRR14596858	KY406782
Stewartia pseudocamellia	MO-6587810	Missouri Botanical Garden	83 044 916	SRR14596847	KY406786
Stewartia pteropetiolata	YXQ038	Yunnan, China	124 092 532	SRR14596838	KY406770
Stewartia rostrata	YXQ15072003	Jiangxi, China	91 741 812	SRR14596837	KY406789
Stewartia rubiginosa	YXQ189	Hunan, China	91 059 484	SRR14596836	KY406777
Stewartia sinensis	YXQ15072001	Jiangxi, China	91 325 600	SRR14596835	KY406748
Gordonieae					
Franklinia alatamaha	MO-6587811	Missouri Botanical Garden	101 387 448	SRR14596890	KY406774
Gordonia brenesii	N. Zamora 7196	Guanacaste, Costa Rica	104 462 244	SRR14596889	KY406761
Gordonia lasianthus	JCRA 110687	JC Raulston Arboretum, Raleigh, NC, USA	82 908 172	SRR14596888	KY406790
Schima argentea	YXQ226	Yunnan, China	86 065 752	SRR14596887	–*
Schima brevipedicellata	YXQ072	Yunnan, China	89 409 700	SRR14596886	KY406784
Schima noronhae	YXQ034	Yunnan, China	88 824 452	SRR14596885	KY406787
Schima sericans	YXQ053	Yunnan, China	97 108 488	SRR14596884	KY406779
Schima superba	YXQ142	Guangxi, China	91 704 880	SRR14596883	KY406788
Schima wallichii	YXQ001	Yunnan, China	89 876 764	SRR14596882	KY406795
Theeae					
Apterosperma oblata	YangSX 5978	Guangdong, China	86 466 584	SRR14596881	–
Camellia amplexifolia	YangSX 5010	Hainan, China	98 672 716	SRR14596879	–
Camellia assimiloides	YangSX 5540	Guangdong, China	98 123 144	SRR14596878	–
Camellia cordifolia	YangSX 5551	Guangdong, China	96 616 940	SRR14596877	–
Camellia cuspidata	YangSX 5118	Hubei, China	78 369 336	SRR14596876	–
Camellia flavida	YangSX 5865	Guangxi, China	92 242 904	SRR14596875	–
Camellia fluviatilis	YangSX 4033	Guangxi, China	68 780 488	SRR14596874	–
Camellia grijsii	YangSX 6064	Guizhou, China	120 959 328	SRR14596873	–
Camellia gymnogyna	YangSX 5953	Guangxi, China	94 768 804	SRR14596872	–
Camellia huana	YangSX 5653	Guizhou, China	111 316 580	SRR14596871	–
Camellia ilicifolia	YangSX 5287	Guizhou, China	95 559 812	SRR14596870	–
Camellia longipedicellata	YangSX 5926	Guangxi, China	90 347 792	SRR14596868	–
Camellia longissima	YangSX 5079	Guangxi, China	99 050 172	SRR14596867	–
Camellia luteoflora	YangSX 6063	Guizhou, China	123 788 308	SRR14596866	–
Camellia pilosperma	YangSX 4714	Guangxi, China	92 105 712	SRR14596865	–
Camellia pitardii var. *compressa*	YangSX 4576	Hunan, China	76 531 640	SRR14596864	–
Camellia semiserrata	YangSX 5555	Guangdong, China	86 751 660	SRR14596863	–
*Camellia sinensis*var. *pubilimba*	YangSX 5927	Guangxi, China	85 785 132	SRR14596862	–
Camellia szechuanensis	YangSX 5064	Sichuan, China	88 726 296	SRR14596861	–
Camellia tsingpienensis	YangSX 5798	Guangxi, China	90 514 420	SRR14596860	–
Camellia tuberculata	YangSX 5202	Chongqing, China	90 076 900	SRR14596859	–
Laplacea fruticosa	NZ10477	Puntarenas, Costa Rica	1 532 313 620	SRR14596857	–
Polyspora axillaris	YXQ099	Hainan, China	92 917 500	SRR14596856	KY406760
Polyspora chrysandra	YXQ221	Yunnan, China	84 011 964	SRR14596855	–
Polyspora dalgleishiana	BROWP 501	Royal Botanic Garden Edinburgh, UK	83 838 300	SRR14596854	KY406769
Polyspora hainanensis	YXQ097	Hainan, China	97 170 208	SRR14596853	KY406776
Polyspora longicarpa	YangSX 4779	Yunnan, China	89 251 900	SRR14596852	KY406768
Polyspora speciosa	YXQ145	Guangxi, China	96 508 612	SRR14596851	KY406754
Pyrenaria hirta var. *cordatula*	YXQ169	Guangxi, China	92 426 012	SRR14596850	KY406785
Pyrenaria hirta var. *hirta*	YangSX 4067	Guangxi, China	100 342 800	SRR14596849	–
Pyrenaria jonquieriana subsp. multisepala	YXQ106	Hainan, China	87 583 392	SRR14596848	KY406772
Pyrenaria khasiana	YangSX 5046	Myanmar, Kachin	88 528 812	SRR14596846	KY406756
Pyrenaria menglaensis	YXQ211	Yunnan, China	87 085 244	SRR14596845	KY406747
Pyrenaria microcarpa var. *microcarpa*	YXQ101	Hainan, China	81 080 788	SRR14596844	KY406764
Pyrenaria oblongicarpa	YXQ216	Yunnan, China	98 658 408	SRR14596843	KY406781
Pyrenaria pingpienensis	YXQ210	Yunnan, China	96 601 500	SRR14596842	–
Pyrenaria shinkoensis	YangSX 5038	Taiwan, China	101 120 720	SRR14596841	–
Pyrenaria spectabilis var. *greeniae*	YXQ172	Hunan, China	93 451 072	SRR14596840	KY406753
Pyrenaria spectabilis var. spectabilis	YXQ155	Guangxi, China	88 012 020	SRR14596839	KY406765

TABLE 1. *Continued*

Taxon	Voucher specimen	Sources	No. of reads (trimmed)	SRA number	Plastid genome
Outgroups					
Galax urceolata	Diapensiaceae	–	9 647 946	ERX2099546	–
Eurya acuminatissima	Pentaphylacaceae	–	25 009 135	SRX2786652	–
Ternstroemia gymnanthera	Pentaphylacaceae	–	14 421 549	ERX2099558	–
Sinojackia xylocarpa	Styracaceae	–	10 611 525	ERX2099565	–
Symplocos tinctoria	Symplocaceae	–	10 196 542	ERX2099566	–
Symplocos paniculata	Symplocaceae	–	28 374 406	SRX1601992	–

*A dash represents the plastid genes were extracted from the transcriptome data or no data were available.

previous studies (Prince and Parks, 2001; S. X. Yang *et al.*, 2004; M. M. Li *et al.*, 2013; Yu *et al.*, 2017*b*), these 21 orthologues probably contain hidden paralogues and could be problematic for downstream analysis. Therefore, we excluded these 21 orthologues to yield a final gene dataset with 610 orthologues (hereafter referred to as the reduced 610 orthologues dataset). Functional annotations for the 610 and 21 orthologues were performed using eggnog-mapper-2.1.4 (Huerta-Cepas *et al.*, 2017) with the command diamond; the best hit was chosen as the final annotation. Annotation results were visualized in WEGO2.0 (https://wego.genomics.cn/).

Sequence assembly for plastid genomes

To construct the plastid matrix, plastid protein-coding genes from 27 samples were extracted from the transcriptome or whole-genome sequencing data generated here, while the plastid genomes from 31 samples (Table 1) were obtained from complete plastid genomes already available from our previous study (Yu *et al.*, 2017*b*). For those species without plastid genome data, we first assembled the plastid genome using GetOrganelle v1.6.2e (Jin *et al.*, 2020), and the protein-coding sequences (CDS) were extracted from the transcriptome and whole-genome sequencing data. For those species with previously completed plastid genomes, protein-coding sequences were extracted following parallel methods to yield a combined matrix of all 58 samples (57 species) for which nuclear gene sequences were obtained.

Phylogenetic analyses based on nuclear genes and plastid genome

For nuclear genes, the obtained nucleotide sequences were aligned with MAFFT v7.407 (Katoh and Standley, 2013). Alignment statistics were calculated by AMAS (Borowiec, 2016). To better assess evolutionary history, both concatenation and coalescent approaches were used to reconstruct intergeneric relationships of Theaceae. For concatenation, partitioned ML and BI analyses were performed using RAxML v8.2.12 (Stamatakis, 2014) and MrBayes v3.2.6 (Ronquist *et al.*, 2012), respectively. Partitioning schemes and models were selected using PartitionFinder v2.1.1 for 610 orthologues (Lanfear *et al.*, 2016), and 188 subsets were obtained for the dataset. In the ML analysis, BS values were calculated using 1000 replicates. In the BI analysis, four chains were run for 2 000 000 generations with random initial trees; every 100 generations, trees were sampled, and the first 25 % of the trees were discarded as burn-in.

For the coalescent analysis, an ML gene tree was reconstructed for each orthologue with RAxML using the same parameter settings as above. The best ML gene tree and 100 bootstrap replicate trees generated for each orthologue set were used to estimate the species tree and supporting values in ASTRAL-III v5.6.3 (C. Zhang *et al.*, 2018); nodes with bootstrap support below 10 % in all gene trees were removed using Newick utilities (Junier and Zdobnov, 2010), in order to improve accuracy in the ASTRAL analysis (C. Zhang *et al.*, 2018). Support of the species tree was quantified using the local posterior probability (LPP) of a branch as a function of its normalized quartet support (Sayyari and Mirarab, 2016). The co-phylogenetic plot was created using the phytools package (Revell, 2012) in R to visualize the differences between trees. We used PhyParts (Smith *et al.*, 2015) to examine patterns of gene tree concordance and conflict within the nuclear genome. All 610 gene trees were rooted using Phyx (Brown *et al.*, 2017), and outgroups were removed. The results were visualized using the program phypartspiecharts (https://github.com/mossmatters/phyloscripts/).

For the plastid genome sequence data, we obtained an 80-CDS dataset for the same taxa represented in the nuclear dataset. The nucleotide sequences were aligned with MAFFT v7.407 (Katoh and Standley, 2013). Phylogenetic reconstructions followed the approaches employed in our previous study (Yu *et al.*, 2017*b*).

Evolutionary network analysis

To test the previously proposed hypothesis of intergeneric hybridization in Theaceae (W. Zhang *et al.*, 2014), we used PhyloNet v3.8.2 (Than *et al.*, 2008) to infer an evolutionary network for Theaceae, using the command 'InferNetwork_MPL' under a maximum pseudo-likelihood framework (Yu and Nakhleh, 2015) and 505 individual gene trees with one outgroup species. Given that the computational time needed by network methods scales very rapidly with taxon number (Folk *et al.*, 2018), we reduced the sampling to 22 species, which is a computationally tractable size (i.e. <30 species; Than *et al.*, 2008; Wen *et al.*, 2018). Given that we mainly focused on hybridization events between genera, representative species from

all nine genera were used. Specifically, we selected one species for each monotypic genus or small genus (i.e. *Apterosperma*, *Franklinia* and *Laplacea*) and selected two (e.g. *Schima*) to six (e.g. *Camellia*) species representing the main subclades of each species-rich genus. This level of taxon sampling might impact our ability to detect reticulations within genera but would not necessarily impact our primary goal of investigating deep inter-generic introgression events. Only one outgroup species was used (different outgroup species were selected for each gene tree because of missing data) (Table 1). The maximum pseudo-likelihood algorithm requires a priori specification of the number of reticulating branches; the number of reticulations was set as one, two, three and four in repeated analyses, as per developer recommendations. Gene trees with branches with <70 % BS were collapsed, and five optimal networks were returned for each analysis. The command CalGTProb was used to compute the likelihood scores and select the best network.

Molecular dating based on 610 nuclear genes

We extracted the variable sites (187 255 bp in length) of the 610 orthologues dataset using MEGA v7 (Kumar *et al.*, 2016) for our molecular dating analysis. Three species from Symplocaceae and Styracaceae, the most closely related families to Theaceae, were used as outgroups. Bayesian estimations of divergence times were conducted in BEAST v2.6.4 (Bouckaert *et al.*, 2014), using the GTR + I + G nucleotide substitution model, lognormal uncorrelated relaxed clock model and birth–death tree prior. We selected two fossil calibration points within Theaceae following our previous study (Table 2; Yu *et al.*, 2017*b*) and conservatively set the two calibration constraints to the stems of those clades. Each fossil age was used to constrain the 'offset' in the lognormal distribution, with the mean 'M' set as 20 % (e.g. C4 in Table 2: offset = 23, M = 4.6) of the fossil age (checked '*Mean in Real Space*') and the standard deviation '*S*' set to 1.0. Additionally, three secondary calibration points were also used following our previous study (Yu *et al.*, 2017*b*) and other studies which also used the Bayesian dating method such as BEAST or MCMCTREE (Magallón *et al.*, 2015; Foster *et al.*, 2017). First, the root of the tree, i.e. the stem age of Theaceae, was constrained under a uniform prior as 79.8–102.5 Mya; 79.8 and 102.5 Mya represent the minimum and maximum age of the 95 % HPD (highest posterior density) of the Theaceae stem in previous studies (Supplementary Data Table S1; Magallón *et al.*, 2015; Foster *et al.*, 2017; Yu *et al.*, 2017*b*). Second, the crown of Theaceae and crown of Gordonieae and Theeae were also constrained under a uniform prior using the ages obtained from our previous study (Table 2; Yu *et al.*, 2017*b*). For each analysis, we ran one billion generations with sampling every 10 000 generations. Convergence was attained within 500–600 million generations, and the effective sample size (ESS) values for all parameters were >100. We removed the first 600 million generations as burn-in and used the remaining 40 000 trees to generate the maximum clade credibility (MCC) tree by TREEANNOTATOR v2.6.4 (with a PP limit of 0.5 and median node heights).

Inference of whole-genome duplication

To investigate the putative ancient WGD in Theaceae, we applied the Python package 'wgd' (Zwaenepoel and Van de Peer, 2019) to construct synonymous substitution (K_S) distributions (ranging from 0.05 to 3) among paralogues from 56 Theaceae transcriptomes and the six outgroup transcriptomes noted in Table 1. Using the command 'mcl' to blast and cluster sequences with each CDS, the commands 'ksd' and 'mix' were used to construct the *Ks* distribution and mixture modelling of *Ks* distributions, respectively. For analysis of the mixture model, we used the BGMM method in the wgd package.

RESULTS

Characteristics of transcriptomes and datasets

We sequenced transcriptomes (ranging from 5.85 to 10.6 Gb) of 56 individuals from 55 species and obtained 138.7 Gb and 165.7 Gb (~89.5× coverage, based on an estimated genome size of 1.85 Gb for *Stewartia pteropetiolata* from flow cytometry) of genomic data for *Laplacea fruticosa* and *Stewartia malacodendron*, respectively. From trimmed reads, assembly of the 64 transcriptomes/genomes (56 transcriptomes and two genomes generated in this study and six transcriptomes from GenBank) provided an average length of unigenes from 382 to 1578 bp. The N50 length ranged from 377 to 1959 bp, with an average of 1491 bp (Supplementary Data Table S1). In total, 610 orthologues were obtained from 56 transcriptomes and two whole-genome sequencing datasets, with the aligned length of the orthologue sets ranging from 309 to 8854 bp and the

TABLE 2. *Fossils used for calibrations in this study, selected from Yu* et al. *(2017)*

Calibration nodes	Fossils	Calibration types	Ages (epoch)	Fossil assignment	Constrained age	References
C1	NA	Secondary calibration	79.8–102.5	Root of the tree (stem of Theaceae)	79.8–102.5	(Foster *et al.*, 2017; Yu *et al.*, 2017*b*)
C2	NA	Secondary calibration	39.6–74.7	Crown of Theaceae	39.6–74.7	(Yu *et al.*, 2017*b*)
C3	NA	Secondary calibration	33.8–66.9	Crown of Gordonieae and Theeae	33.8–66.9	(Yu *et al.*, 2017*b*)
C4	*Schima kwangsiensis* X. G. Shi, C. Quan et J. H. Jin	Fruits, seeds fossil	Late Oligocene	Stem of *Schima*	23.0	(M. M. Li *et al.*, 2013; Quan *et al.*, 2016; X. G. Shi *et al.*, 2017)
C5	*Hartia quinqueangularis*(Menzel)	Fruits, seeds fossil	Late Miocene	Stem of *Hartia* (now a clade within *Stewartia*)	5.3	(Mai, 1975)

proportion of missing data ranging from 0 to 29.65 % (Table S2). The aligned length of the concatenated 610 orthologues was 858 606 bp, with 227 708 (26.5 %) variable sites and 21.67 % missing data. The alignment length of the concatenated 80 plastid coding genes was 69 225 bp, with 7236 (10.5 %) variable sites and 11.88 % missing data. Gene function of the 610 and 21 orthologues did not show a significant difference, with genes from both sets mainly related to the cellular component, molecular function and general biological process (Table S3, Figs S1 and S2). All transcriptomic raw reads have been deposited in GenBank (Table 1), and all of the alignments and trees in this study have been submitted to TreeBASE (see Supplementary Data statement at the end of the paper).

Phylogenetic relationships and networks in Theaceae

The topology recovered from RAxML analyses based on the concatenated 610 low-copy nuclear genes strongly supported a sister relationship between Theeae and Gordonieae (MLBS = 100 %, PP = 1.00; Fig. 1). Additionally, PhyParts analysis indicated 435 out of 610 gene trees (71.3 %) supported this topology (Fig. 2). Our coalescent-based species tree inferred from ASTRAL revealed the same relationship among the three tribes as obtained with the concatenation analysis (LPP = 1.00; Fig. 1; Supplementary Data Fig. S3). RAxML analyses based on the 80 protein-coding genes of the plastid genome also supported a sister relationship between Theeae and Gordonieae (MLBS = 91 %, PP = 1.00; Fig. S4). The co-phylogenetic plot between the concatenated and coalescent tree topology of the 610 nuclear genes (Fig. 1), and between the nuclear coalescent tree and plastome topology (Fig. S4) did not show any incongruence of relationships among the three tribes. Hence, support at the tribal level of Theaceae was strong and uniform across the data partitions and analytical methods employed here.

Generic-level relationships recovered within Gordonieae and Stewartieae were largely consistent with previous phylogenetic studies of Theaceae. Within Theeae, the sister relationships between *Camellia* and *Polyspora* and between *Apterosperma* and *Laplacea fruticosa* were both maximumly supported in all of the analyses using transcriptomic data (Fig. 1; Supplementary Data Fig. S3). Nevertheless, the position of the *Apterosperma–Laplacea* clade varied among analyses. Using the 610 low-copy nuclear genes, the concatenation analysis placed the *Apterosperma–Laplacea* clade grouped with *Pyrenaria* with moderate support (MLBS = 72 %, PP = 1.00; Fig. 1). The ASTRAL topology (Fig. S3), while largely congruent overall with the results from the concatenated analyses (Fig. 1), instead placed *Apterosperma–Laplacea* as sister to the *Camellia–Polyspora* clade with very weak support (LPP = 0.35; Fig. S3). At lower taxonomic levels, PhyParts recovered strong discordance across many parts of the tree, with only 112 out of 610 gene trees supporting the sister relationship between the *Apterosperma–Laplacea* clade and the *Camellia–Polyspora* clade, and 488 gene trees supporting conflicting/alternative resolutions (Fig. 2). Phylogenetic trees based on the dataset of 80 plastid protein-coding genes were highly consistent with our previous study (Yu *et al.*, 2017*b*) with the exception that *Apterosperma* was sister to *Camellia* (MLBS = 86 %, PP = 1.00; Fig. S4).

For the PhyloNet analyses, the inferred network with the highest log pseudo-likelihood (−8792.9795) included three reticulation events (reticulation numbers were set as 4; Fig. 3). This analysis suggested one reticulation between *Camellia tsingpienensis* and *C. amplexifolia*, and the descendant putative hybrid further backcrossed with *C. tsingpienensis* and formed the clade comprising *C. huana* and *C. szechuanensis*. We also recovered evidence of a reticulation event suggesting that *Franklinia alatamaha* descended from a putative ancient hybridization event between the common ancestor of *Gordonia* and the common ancestor of *Schima*. All other analyses (reticulation numbers = 1, 2, 3), for which likelihood was suboptimal, only detected reticulation within *Camellia* (Fig. 3).

Divergence time estimation

The stem and crown ages of Theaceae were estimated to be 99.7 Mya (95 % HPD: 92.0–102.5) and 64.9 Mya (56.4–73.3), respectively (Fig. 4; Supplementary Data Table S4). The divergence between Gordonieae and Theeae was ~55.7 Mya (48.1–62.8). The crown ages of Stewartieae, Gordonieae and Theeae were estimated to be 18.5 Mya (15.6–21.7), 25.1 Mya (24.1–26.6) and 22.1 Mya (19.2–25.2), respectively. The estimated ages were all within the 95 % HPD of our previous study (Yu *et al.*, 2017*b*) and Yan *et al.* (2021).

Whole-genome duplication

Ks (synonymous substitution rate) analyses using all 57 ingroup taxa suggested that all species showed a peak at around *Ks* = 1.5, and all except one species from Theaceae presented a peak at around *Ks* = 0.4 (Fig. 4; Supplementary Data Fig. S5). The two peaks are consistent with the two WGD events (At-γ: *Ks* = 1.16; Ad-β: *Ks* = 0.36) reported in the tea (*Camellia sinensis* var. *sinensis*) genome (Xia *et al.*, 2017). No *Ks* peak around 0.4 was found in *Stewartia ovata*; given the nested position of this species, the absence may be an artefact due to stochastic error or may be because the transcriptome data only represent the expressed genes in those tissues (e.g. roots, leaves, flowers) sampled. For the six outgroup species, consistent *Ks* peaks were also found at around 1.5 and 0.4 (Fig. S5), indicating that Symplocaceae, Styracaceae, Pentaphylacaceae and Diapensiaceae shared the two WGD events (i.e. At-γ, Ad-β) with Theaceae.

DISCUSSION

Consistent relationships among tribes between plastome and transcriptome trees

Numerous phylogenetic studies have not been able to elucidate with strong support the relationships among the three tribes within Theaceae, due in part to incomplete taxon sampling and few loci, either a few plastid and nuclear loci or half of the SSC region of the plastid genome (Prince and Parks, 2001; S. X. Yang *et al.*, 2004; J. B. Yang *et al.*, 2006; M. M. Li *et al.*, 2013). Using 25 loci from the plastid, nuclear and mitochondrial

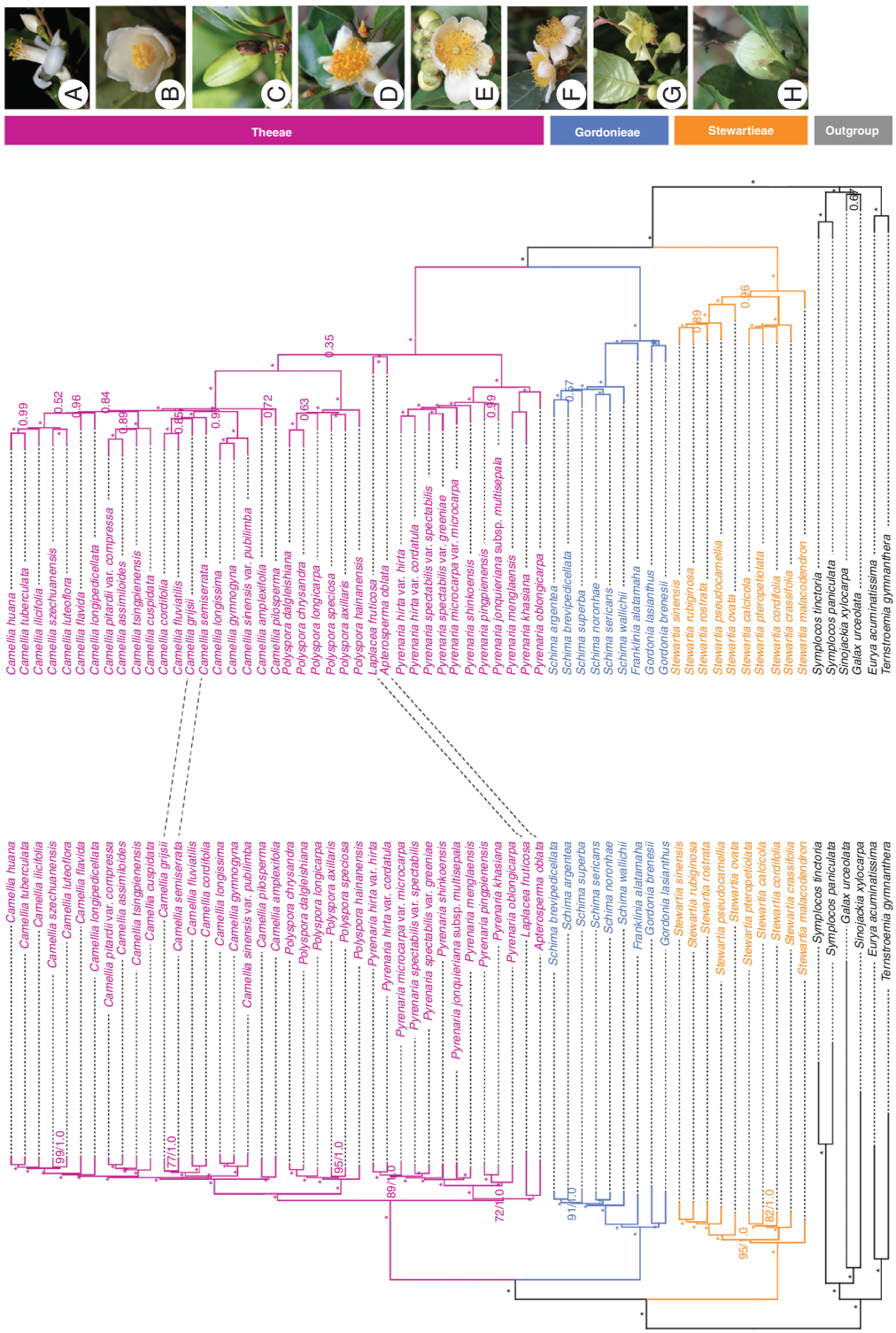

FIG. 1. Co-phylogenetic comparison between the RAxML topology based on a concatenated analysis of 610 nuclear genes (left) and ASTRAL-III nuclear coalescent species tree (right) for Theaceae. Numbers associated with nodes indicate ML bootstrap support (BS)/Bayesian inference (BI) posterior probability (PP) values in the concatenated tree and local posterior probability (LPP) in the species tree. Scale bar denotes the number of substitutions per site. Asterisks represent nodes with maximal support. Branch lengths in the species tree are labelled in coalescent units. Images: a, *Camellia tsingpienensis*; b, *Camellia cordifolia*; c, *Polyspora speciosa*; d, *Pyrenaria hirta* var. *cordatula*; e, *Schima argentea*; f, *Schima wallichii*; g, *Stewartia calcicola*; h, *Stewartia rubiginosa*.

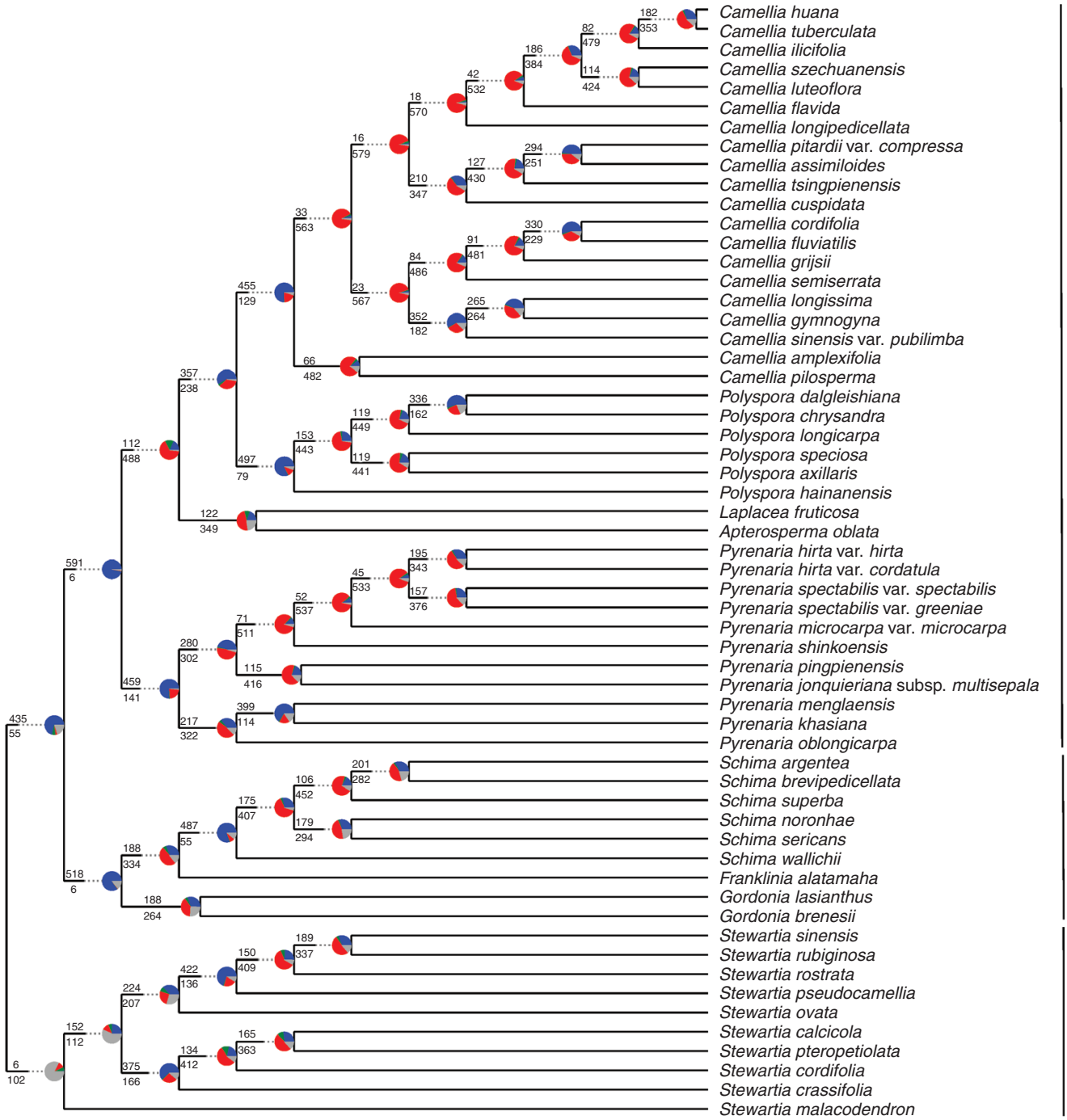

FIG. 2. Patterns of gene-tree concordance and conflict within Theaceae based on the PhyParts analysis. The tree topology used is that inferred by ASTRAL-III. The pie charts at each node show the proportion of genes in concordance (blue), conflict (green = a single dominant alternative; red = all other conflicting trees) and without enough information (grey). The numbers above and below each branch are the numbers of concordant and conflicting genes at each bipartition, respectively.

genomes from 4531 species from Ericales, a recent study found that Theeae and Gordonieae grouped together (MLBS > 70 %, PP > 0.95) with Stewartieae as their sister, but only ten species of Theaceae were included (Rose *et al.*, 2018). Our previous study likewise strongly supported a sister relationship between Theeae and Gordonieae (MLBS = 91 %, PP = 1.00) based on a combined plastome and nuclear ribosomal DNA dataset (Yu *et al.*, 2017*b*). Here we present a phylogenetic framework of Theaceae using 610 orthologous low-copy nuclear genes. The topology obtained from both the concatenation and the coalescence analyses of the 610 low-copy nuclear genes consistently and strongly supported a sister relationship between Theeae

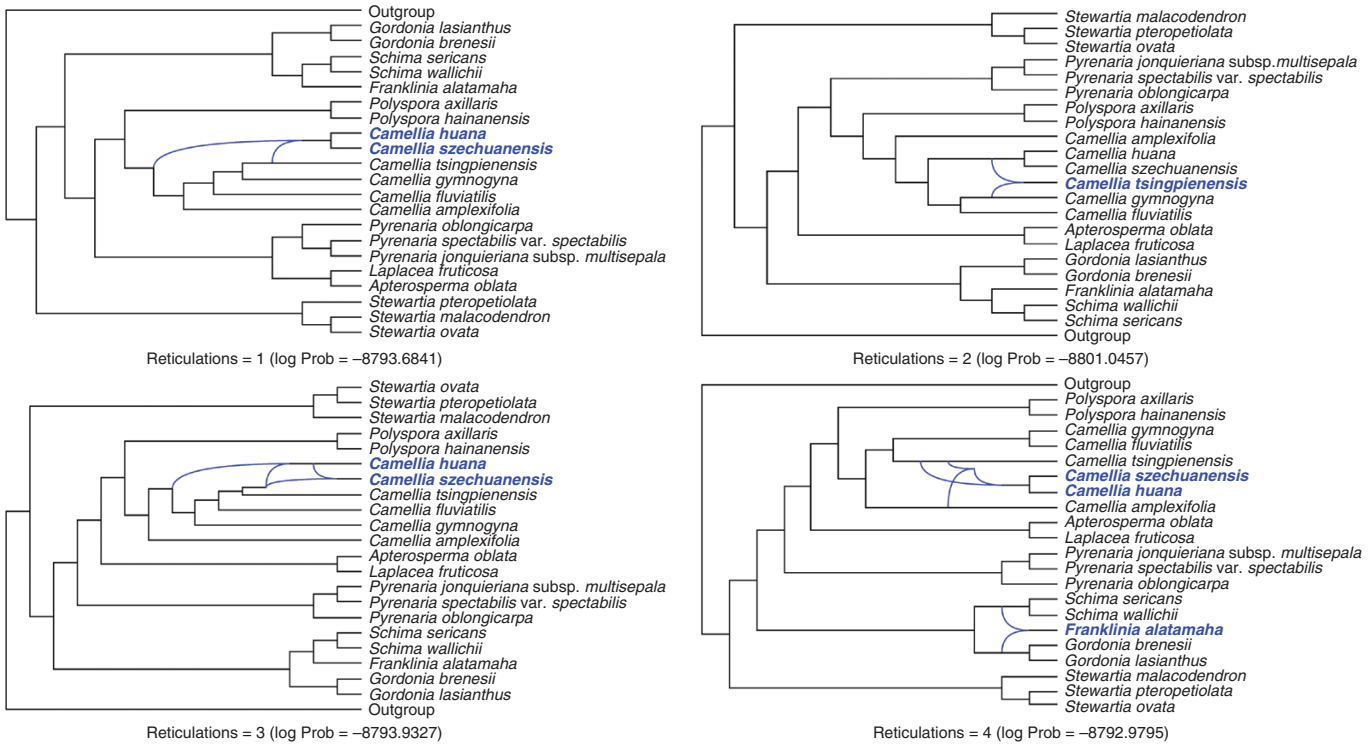

FIG. 3. The optimal phylogenetic network of Theaceae inferred using PhyloNet, with the number of reticulations as 1, 2, 3 and 4. The scenario with four reticulations had the best log pseudo-likelihood.

and Gordonieae (MLBS = 100 %, PP = 1.00, LPP = 1.00; Fig. 1; Supplementary Data Fig. S3), consistent with the results from the plastid genome (MLBS = 91 %, PP = 1.00; Fig. S4). In addition, 435 of the 610 low-copy nuclear genes supported this topology (Fig. 2). The ((Gordonieae, Theeae) Stewartieae) relationship is consistent with the evolutionary pattern of the endosperm in Theaceae, as discussed in Yu *et al.* (2017*b*).

Improved intergeneric relationships based on transcriptome data

Intergeneric relationships within Gordonieae and Stewartieae have been fully resolved in previous studies (Prince and Parks, 2001; S. X. Yang *et al.*, 2004; J. B. Yang *et al.*, 2006; M. M. Li *et al.*, 2013; Yu *et al.*, 2017*b*). However, the relationships among the five genera in Theeae have been controversial. *Laplacea* was placed in a clade comprising *Camellia*, *Tutcheria* (=*Pyrenaria*) and *Glyptocarpa* (=*Camellia*) (Prince and Parks, 2001). W. Zhang *et al.* (2014) revealed that *Apterosperma* formed a sister relationship with *Polyspora* (MLBS = 73 %, PP = 1.00) in the plastid DNA tree, but these two genera were placed in a clade comprising *Tutcheria* (=*Pyrenaria*) and *Parapyrenaria* (=*Pyrenaria*) (MLBS = 68 %, PP = 0.72) in the *LEAFY* tree. Based on the 610 low-copy nuclear genes, the resolution of the relationships among the five genera in Theeae has been improved. The *Apterosperma–Laplacea* clade received maximal support in the concatenation analyses using the 610 low-copy nuclear genes and grouped with *Pyrenaria* with moderate support (MLBS = 72 %, PP = 1.00; Fig. 1). However, the ASTRAL topology suggested that the *Apterosperma–Laplacea* clade was weakly supported as sister to the *Camellia–Polyspora* clade (LPP = 0.35; Fig. 1; Supplementary Data Fig. S3); 112 out of 610 nuclear genes supported this topology (Fig. 2). The strongly supported *Apterosperma–Laplacea* clade also grouped with *Pyrenaria* with moderate support based on the whole plastid genome dataset (MLBS = 67 %), the SSC (MLBS = 80 %) dataset and the protein-coding gene dataset (MLBS = 75 %) from our previous study (Yu *et al.*, 2017*b*). Taken as a whole, based on evidence from both plastid genomes and transcriptome data, we suggest (((*Apterosperma–Laplacea*), *Pyrenaria*), (*Camellia–Polyspora*)) as the most likely topology.

Phylogenetic network inference suggests three reticulation events in Theaceae

In W. Zhang *et al.* (2014), *Camellia* and *Pyrenaria* were not recovered as monophyletic, and widespread hybridization among genera in Theeae was proposed. Phylogenetic conflict found in *Stewartia* was also suggested to be caused by ancient introgressive hybridization following species diversification, leading to discordant histories in the nuclear and plastid genomes (H. Y. Lin *et al.*, 2019). However, while our PhyloNet analyses supported the presence of hybridization in the history of Theaceae (Fig. 3), they do not support the specific reticulation scenario suggested by W. Zhang *et al.* (2014); *Camellia* and *Pyrenaria* were supported as monophyletic based on the 610 low-copy nuclear genes (Fig. 1; Supplementary Data Fig. S3). Two of the three reticulation events detected in the best-fit network were within *Camellia*, and another intergeneric reticulation was in Gordonieae, but not Theeae (Fig. 3).

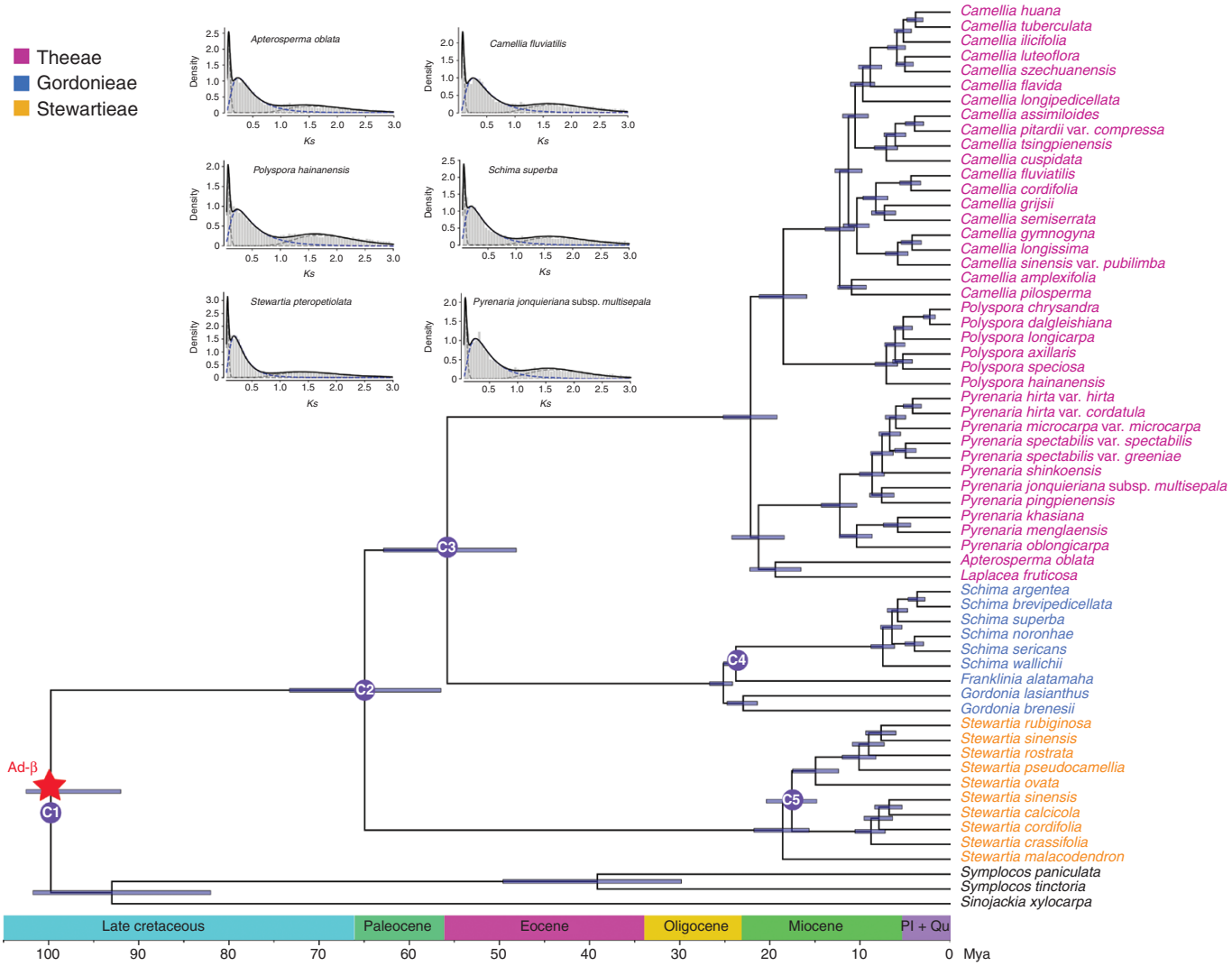

FIG. 4. Chronogram of Theaceae derived from BEAST analysis using 610 nuclear genes with *Ks* distribution plots for paralogues showing inferred WGDs (blue dashed line) in the upper left corner; the six species included represent all three tribes of Theaceae. The ages of stratigraphic boundaries were obtained from the International Chronostratigraphic Chart (Cohen *et al.*, 2013) (Pl, Pliocene; Qu, Quaternary), with a scale as millions of years ago (Mya). Blue bars at each node represent the 95 % highest posterior density (HPD) with posterior probability above 0.5. C1–C5 represent the five calibration points used, which correspond to those nodes listed in Table 2. The red star represents the WGD event shared by all members of the clade including core Ericales, primuloids, polemonioids and Lecythidaceae, reported in Zhang *et al.* (2020).

The estimated stem and crown ages of Theaceae are largely consistent with previous studies (M. M. Li *et al.*, 2013; Magallón *et al.*, 2015; Foster *et al.*, 2017; Yu *et al.*, 2017*b*; Rose *et al.*, 2018; Yan *et al.*, 2021) (Supplementary Data Table S4). Our study supported the findings of Yan *et al.* (2021) that Theaceae are of boreotropical forest origin. Under the deep reticulation scenario (Fig. 3), *Franklinia alatamaha* (the only extant species of the genus, distributed in eastern North America) descends from a putative ancient hybridization event between the common ancestor of *Gordonia* (endemic to North and Central America) and the common ancestor of *Schima* (endemic to Asia). The molecular dating analysis suggests a date of reticulation during the Late Oligocene, ~25.1 Mya (24.1–26.6) (Fig. 4; Table S4). A recent study also suggested that gene flow between species of Theaceae from different continents (e.g. North America, Eurasia) must have occurred during the Oligocene and possibly until the mid-Miocene (Yan *et al.*, 2021). In the study by Yan *et al.* (2021), the ancestral distribution of the crown of Gordonieae was inferred as Nearctic + Sino-Japanese (NS); it is likely that the ancestors of *Gordonia* and *Schima* co-occurred across the NS region during the late Oligocene (~25.6 Mya), and hence ancient gene flow between them was possible. This is a plausible scenario as land bridges (e.g. the Bering land bridge) existed during the Late Oligocene, and the eastern Asia and eastern North American flora was probably continuous across high latitudes of the Northern Hemisphere (Tiffney, 1985; Tiffney and Manchester, 2001; Milne, 2006), allowing for species contact and opportunities for hybridization, which is no longer possible under today's

climate and geography. Fossils of Theaceae are known from mid-latitude Northern Hemisphere localities such as western Kentucky and Tennessee in North America and Germany in Europe during the Eocene and Oligocene (Grote and Dilcher, 1992; Kvacek and Walther, 1998; Wilde and Frankenhauser, 1998; Kvaček, 2004).

Overall, our work supports ancient introgression within Theaceae, and is consistent with biogeographical patterns and the fossil history of the group, adding to an increasing list of ancestral hybrids among currently allopatric taxa that yield unique evidence of past biogeographical distributions (e.g. Folk *et al.*, 2018). Interspecific gene flow of an intercontinental scope, although perhaps less geographically remarkable considering that the past distribution of these plants was probably higher in latitude, has likewise been reported in other plant groups distributed across Eurasia and North America. For example, phylogenetic and fossil evidence supports a North American origin for *Picea* (Pinaceae); the Bering land bridge may have facilitated the introgression between species from North America and Eurasia during the Miocene and Pliocene (Ran *et al.*, 2015). Discordance of the mtDNA tree with nuclear and cpDNA trees of *Abies* (Pinaceae) also indicated intercontinental migration and introgressive hybridizations in this genus during the Miocene (Semerikova *et al.*, 2018). More detailed studies are needed to explore the degree and frequency of intercontinental gene flow between Eastern Asia and North America in other lineages. In particular, a continued recovery of a primarily Miocene date for past hybridization would be of interest in understanding when plant biogeographical connections ceased among these areas.

The best-fit network suggested the clade comprising *C. huana* and *C. szechuanensis* was formed through two rounds of hybridizations, with *Camellia tsingpienensis* and *C. amplexifolia* as parents. All other PhyloNet analyses (reticulations = 1, 2, 3) consistently indicated a clear pattern of intrageneric gene flow within *Camellia* (Fig. 3). *Camellia tsingpienensis* is found in Guangxi, south-eastern Yunnan of China and northern Vietnam, and *C. amplexifolia* is only present in Hainan of China. Recent studies have suggested that the flora of Hainan is of continental origin and has the highest floristic affinity with Vietnam, and periodic emergence of land bridges between Hainan and Vietnam during Quaternary glacial cycles might have resulted in their floristic affinity (Ali, 2018; S. L. Lin *et al.*, 2021). Even though there are no distribution overlaps between *C. tsingpienensis* and *C. amplexifolia*, interspecific gene flow was possible when Hainan and the neighbouring landmasses including Vietnam were connected during the glacial periods of the Quaternary.

Previous studies have likewise found evidence of hybridization in *Camellia*. First, cultivated ornamental camellias resulting from hybridization have been widely used in horticulture (Nishimoto *et al.*, 2003; Tanaka *et al.*, 2005; Xu *et al.*, 2018). Second, Cambod tea (cultivated tea of *C. sinensis* var. *assamica*) was suggested to have originated through hybridization between different tea types (Meegahakumbura *et al.*, 2016; Meegahakumbura *et al.*, 2018). Introgression was also detected between the cultivated *C. sinensis* var. *assamica* and *C. taliensis*, with the latter possibly genetically involved in the domestication of *C. sinensis* var. *assamica* (Li *et al.*, 2015). Given our taxon sampling decisions, further work with increased taxon sampling will be needed to uncover further introgression patterns within *Camellia*.

No Theaceae-specific whole-genome duplication event

Two ancient WGD events, namely At-γ and Ad-β, have been identified in the tea plant (*C. sinensis* var. *assamica*) genome (Xia *et al.*, 2017, 2020). However, analysis of genic collinearity reveals that a recent WGD event occurred after the divergence of the tea and kiwifruit lineages, based on the genome of another variety of tea (*C. sinensis* var. *sinensis*) (Wei *et al.*, 2018); Larson *et al.* (2020) later named this WGD Cm-α. Based on a chromosome-scale genome assembly of *C. sinensis* var. *sinensis*, the authors suggested one recent *Camellia* tetraploidization event occurred after the divergence of *C. sinensis* and *Actinidia chinensis* from their common ancestor (J. D. Chen *et al.*, 2020), but the time of the *Camellia* tetraploidization event (58.9–61.7 Mya) was very close to the divergence time between *C. sinensis* and *A. chinensis* at 61.2–65.3 Mya. Here, we identified two WGD events shared by all genera of Theaceae and also representatives of other related families (Symplocaceae, Styracaceae, Pentaphylacaceae and Diapensiaceae; Fig. 4; Supplementary Data Fig. S5). We have clarified that Cm-α proposed by Wei *et al.* (2018) and Larson *et al.* (2020) and the more recent tetraploidization event found by J. D. Chen *et al.* (2020) were actually Ad-β, which has been recently revised to characterize the core Ericales (ACCHβ + DIOSα) according to genome collinearity and also analyses based on the MAPS pipeline (Leebens-Mack *et al.*, 2019), and more specifically to the clade including core Ericales, primuloids, polemonioids and Lecythidaceae, using deep asterid phylotranscriptomic analyses (C. Zhang *et al.*, 2020). Thus, our results support the hypothesis that the tea family experienced two WGD events (i.e. At-γ and Ad-β) in its evolutionary history, with both shared by other families. There is no evidence for any Theaceae-specific WGD event. Thus, this study sheds light on the significance of broad phylogenetic sampling for inferring the number and placement of WGD events.

SUPPLEMENTARY DATA

Supplementary data are available online at https://academic.oup.com/aob and consist of the following. Figure S1: Gene annotation of 610 orthologous low-copy nuclear genes. Figure S2: Gene annotation of 21 putative paralogous low-copy nuclear genes. Figure S3: Species tree topology, inferred by ASTRAL, from the 610 low-copy nuclear genes; posterior probability values are shown beside the nodes. Figure S4: Co-phylogenetic comparison between ASTRAL-III nuclear coalescent species tree and the RAxML plastome topology of Theaceae. Figure S5: Histograms of the distribution of gene duplications with mixture models of inferred WGDs for all sampled ingroup and outgroup species. Table S1: Statistics of sampled transcriptomes. Table S2: Results of AMAS summaries on 610 low-copy nuclear genes. Table S3: Functional annotation of 610 and 21 low-copy nuclear genes. Table S4: Age estimates of Theaceae (millions of years ago) for selected nodes and comparison with results from other studies.

ACKNOWLEDGEMENTS

We thank Prof. De-Zhu Li (Kunming Institute of Botany, Chinese Academy of Sciences) for his great help during the whole process of this study. Thanks to Prof. Liang Fang (Jiujiang University), Drs Jie Cai, Ting Zhang and Ji-Dong Ya for their help with sample collection, and Prof. Hong-Tao Li, Drs Gregory W. Stull, Si-Yun Chen, Yun-Long Liu, Xin-Yu Du, Fei Zhao, Yin-Zi Jiang, Zhi-Qiong Mo (Kunming Institute of Botany, Chinese Academy of Sciences), Chao Feng (South China Botanical Garden, Chinese Academy of Sciences), Cai-Fei Zhang (Wuhan Botanical Garden, Chinese Academy of Sciences) and Tai-Kui Zhang (Fudan University) for their help with data analysis, and Drs Mark Whitten (our late friend) and Sheng-Chen Shan (University of Florida) for their assistance in sampling *Stewartia malacodendron*.

FUNDING

This work was supported by National Natural Science Foundation of China (Nos. 32070369, 31700182), the Large-scale Scientific Facilities of the Chinese Academy of Sciences (No. 2017-LSFGBOWS-02), the Open Research Fund of Guangxi Key Laboratory of Special Non-wood Forest Cultivation & Utilization (No. 19-B-01-03), the Youth Innovation Promotion Association CAS (No. 2021393), the CAS 'Light of West China' Program and US National Science Foundation Grant (No. DEB-1442280) to Pamela S. Soltis, Douglas E. Soltis and others.

LITERATURE CITED

Ali JR. 2018. New explanation for elements of Hainan Island's biological assemblage may stretch things a little too far. *Ecography* **41**: 457–460.

APG IV. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society* **181**: 1–20.

Barker MS, Kane NC, Matvienko M, *et al.* 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Molecular Biology and Evolution* **25**: 2445–2455.

Bentham G, Hooker J. 1862. *Genera plantarum*. London: Reeve.

Borowiec ML. 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ* **4**: e1660.

Bouckaert R, Heled J, Kühnert D, *et al.* 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* **10**: e1003537.

Brown JW, Walker JF, Smith SA. 2017. Phyx: phylogenetic tools for unix. *Bioinformatics* **33**: 1886–1888.

Calvo J, Alvarez I, Aedo C, Pelser PB. 2013. A phylogenetic analysis and new delimitation of *Senecio* sect. *Crociseris* (Compositae: Senecioneae), with evidence of intergeneric hybridization. *Taxon* **62**: 127–140.

Chang HD, Ren SX 1998. Theaceae (1). In: Wu CY, ed. *Flora Reipublicae Popularis Sinicae*. Beijing: Science Press, 1–251.

Chang HT. 1976. *Apterosperma*-genus novum Theacearum. *Acta Scientiarum Naturalium Universitatis Sunyatseni* **15**: 90–92.

Chen JD, Zheng C, Ma JQ, *et al.* 2020. The chromosome-scale genome reveals the evolution and diversification after the recent tetraploidization event in tea plant. *Horticulture Research* **7**: 63.

Chen SF, Zhou YQ, Chen YR, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**: 884–890.

Cohen KM, Finney SC, Gibbard PL, Fan JX. 2013. The ICS International chronostratigraphic chart. *Episodes* **36**: 199–204.

Davidson NM, Oshlack A. 2014. Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. *Genome Biology* **15**: art410.

Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical bulletin* **19**: 11–15.

Folk RA, Mandel JR, Freudenstein JV. 2017. Ancestral gene flow and parallel organellar genome capture result in extreme phylogenomic discord in a lineage of angiosperms. *Systematic Biology* **66**: 320–337.

Folk RA, Soltis PS, Soltis DE, Guralnick R. 2018. New prospects in the detection and comparative analysis of hybridization in the tree of life. *American Journal of Botany* **105**: 364–375.

Foster CSP, Sauquet H, Merwe Mv, *et al.* 2017. Evaluating the impact of genomic data and priors on Bayesian estimates of the angiosperm evolutionary timescale. *Systematic Biology* **66**: 338–351.

Grabherr MG, Haas BJ, Yassour M, *et al.* 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**: 644–652.

Grayum MH, Madrigal QJ. 2011. A new combination in *Gordonia* (Theaceae). *Phytoneuron* **10**: 1–3.

Grote PJ, Dilcher DL. 1992. Fruits and seeds of tribe Gordonieae (Theaceae) from the Eocene of North America. *American Journal of Botany* **79**: 744–753.

Guo X, Thomas D, Saunders R. 2018. Gene tree discordance and coalescent methods support ancient intergeneric hybridisation between *Dasymaschalon* and *Friesodielsia* (Annonaceae). *Molecular Phylogenetics and Evolution* **127**: 14–29.

Haas BJ, Papanicolaou A, Yassour M, *et al.* 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* **8**: 1494–1512.

Haas BJ, Salzberg SL, Zhu W, *et al.* 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biology* **9**: R7.

Huang CH, Sun R, Hu Y, *et al.* 2015. Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Molecular Biology and Evolution* **33**: 394–412.

Huang CH, Zhang CF, Liu M, *et al.* 2016. Multiple polyploidization events across Asteraceae with two nested events in the early history revealed by nuclear phylogenomics. *Molecular Biology and Evolution* **33**: 2820–2835.

Huang SX, Ding J, Deng DJ, *et al.* 2013. Draft genome of the kiwifruit *Actinidia chinensis*. *Nature Communications* **4**: 2640.

Huerta-Cepas J, Forslund K, Coelho LP, *et al.* 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Molecular Biology and Evolution* **34**: 2115–2122.

IUCN. 2020. *The IUCN Red List of Threatened Species, Version 2020-1*. https://www.iucn.org

Jiao YN, Wickett NJ, Ayyampalayam S, *et al.* 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**: 97–100.

Jin JJ, Yu WB, Yang JB, *et al.* 2020. GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biology* **21**: 241.

Junier T, Zdobnov EM. 2010. The Newick utilities: high-throughput phylogenetic tree processing in the Unix shell. *Bioinformatics* **26**: 1669–1670.

Kajitani R, Toshimoto K, Noguchi H, *et al.* 2014. Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research* **24**: 1384–1395.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**: 772–780.

Kobuski CE. 1949. Studies in the Theaceae XVIII. The West Indian species of *Laplacea*. *Journal of the Arnold Arboretum* **30**: 166–186.

Kobuski CE. 1950. Studies in the Theaceae. XX. Notes on the South and Central American species of *Laplacea*. *Journal of the Arnold Arboretum* **31**: 405–429.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* **33**: 1870–1874.

Kvaček Z. 2004. Revisions to the early Oligocene flora of Flörsheim (Mainz Basin, Germany) based on epidermal anatomy. *Senckenbergiana Lethaea* **84**: 1–73.

Kvacek Z, Walther H. 1998. The Oligocene Volcanic Flora of Kundratice near Litomerice, Ceske Stredohori volcanic complex (Czech Republic) – a review. *Sbornik Narodniho Muzea v Praze, Rada B - Prirodni Vedy (Acta Musei Nationalis Pragae, Series B, Historia Naturalis)* **54**: 1–42.

Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. 2016. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution* **34**: 772–773.

Larson DA, Walker JF, Vargas OM, Smith SA. 2020. A consensus phylogenomic approach highlights paleopolyploid and rapid radiation in the history of Ericales. *American Journal of Botany* **107**: 7731–7789.

Lee SL, Yang TYA. 2019. *Camellia chinmeii*, a new species of *Camellia* sect. *Paracamellia* in Taiwan. *Taiwania* **64**: 321–325.

Leebens-Mack JH, Barker MS, Carpenter EJ, *et al.* 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**: 679–685.

Li MM, Li JH, Del Tredici P, Corajod J, Fu CX. 2013. Phylogenetics and biogeography of Theaceae based on sequences of plastid genes. *Journal of Systematics and Evolution* **51**: 396–404.

Li MM, Meegahakumbura Kasun M, Yan LJ, Liu J, Gao LM. 2015. Genetic involvement of *Camellia taliensis* in the domestication of *C. sinensis* var. *assamica* (Assimica Tea) revealed by nuclear microsatellite markers. *Plant Diversity and Resourcues* **37**: 29–37.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658–1659.

Li Y, Awasthi N, Yang J, Li CS. 2013. Fruits of *Schima* (Theaceae) and seeds of *Toddalia* (Rutaceae) from the Miocene of Yunnan Province, China. *Review of Palaeobotany and Palynology* **193**: 119–127.

Lin HY, Hao YJ, Li JH, *et al.* 2019. Phylogenomic conflict resulting from ancient introgression following species diversification in *Stewartia* s.l. (Theaceae). *Molecular Phylogenetics and Evolution* **135**: 1–11.

Lin SL, Chen L, Peng WX, Yu JH, He JK, Jiang HS. 2021. Temperature and historical land connectivity jointly shape the floristic relationship between Hainan Island and the neighbouring landmasses. *Science of the Total Environment* **769**: 144629.

Liu ZW, Chai SF, Wu FY, *et al.* 2020*a*. *Camellia rostrata*, a new species of yellow camellias from Southwest China. *Phytotaxa* **459**: 61–68.

Liu ZW, Fang W, Liu ED, Zhao M, He YF, Yang SX. 2019. *Camellia mingii*, a new species of yellow camellias from Southeast Yunnan, China. *Phytotaxa* **393**: 47–56.

Liu ZW, Ye PM, Li ZH, *et al.* 2020*b*. *Camellia zhaiana* (sect. *Longipedicellata*), a new species of Theaceae from Guangxi, China. *Phytotaxa* **460**: 225–229.

Magallón S, Gómez-Acevedo S, Sánchez-Reyes LL, Hernández-Hernández T. 2015. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytologist* **207**: 437–453.

Mai DH. 1975. Über Früchte und Samen von *Hartia* Dunn (Theaceae). *Friedrich Schiller University Jena Naturwiss* **24**: 463–476.

Mao KS, Ruhsam M, Ma YZ, *et al.* 2019. A transcriptome-based resolution for a key taxonomic controversy in Cupressaceae. *Annals of Botany* **123**: 153–167.

Matasci N, Hung LH, Yan ZX, *et al.* 2014. Data access for the 1,000 Plants (1KP) project. *GigaScience* **3**: 17.

Meegahakumbura MK, Wambulwa MC, Li MM, *et al.* 2018. Domestication origin and breeding history of the tea plant (*Camellia sinensis*) in China and India based on nuclear microsatellites and cpDNA sequence data. *Frontiers in Plant Science* **8**: 2270.

Meegahakumbura MK, Wambulwa MC, Thapa KK, *et al.* 2016. Indications for three independent domestication events for the Tea Plant (*Camellia sinensis* (L.) O. Kuntze) and new insights into the origin of tea germplasm in China and India revealed by nuclear microsatellites. *PLoS One* **11**: e0155369.

Melchior H 1925. Theaceae. In: Engle A, Prantl E, eds. *Die naturlichen Pflanzenfamilien*, 2nd edn. Leipzig: Wilhelm Engelmann, 109–154.

Milne RI. 2006. Northern Hemisphere plant disjunctions: a window on Tertiary land bridges and climate change? *Annals of Botany* **98**: 465–472.

Ming TL, Bartholomew B. 2007. Theaceae. In: Wu CY, Raven PH, eds. *Flora of China*. Beijing and Saint Louis: Science Press and Missouri Botanical Garden Press, 366–478.

Mirbel CFB. 1813. Notes pour servir a l'histoire naturelle de la famille des Orangers de M. A.-L. de Jussieu. *Nouveau Bulletin des Sciences par la Société Philomatique* **75**: 382.

Morales-Briones DF, Liston A, Tank DC. 2018. Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus *Lachemilla* (Rosaceae). *New Phytologist* **218**: 1668–1684.

Nishimoto S, Shimizu K, Hashimoto F, Sakata Y. 2003. Interspecific hybrids of *Camellia chrysantha* x *C. japonica* by ovule culture. *Journal of the Japanese Society for Horticultural Science* **72**: 236–242.

Orel G. 2006. A new species of *Camellia* section *Piquetia* (Theaceae) from Vietnam. *Novon* **16**: 244–247.

Orel G, Curry AS. 2015. *In pursuit of hidden camellias: 32 new camellia species from Vietnam and China*. Sydney: Theaceae Exploration Associates.

Orel G, Curry AS. 2019. *Camellia: work in progress*. Sydney: Theaceae Exploration Associates.

Orel G, Wilson PG. 2010*a*. *Camellia luteocerata* sp. nov. and a new section of *Camellia* (Dalatia) from Vietnam. *Nordic Journal of Botany* **28**: 280–284.

Orel G, Wilson PG. 2010*b*. A new species of *Camellia* Sect. *Stereocarpus* (Theaceae) from Vietnam. *Novon* **20**: 198–202.

Orel G, Wilson PG. 2012. *Camellia cherryana* (Theaceae), a new species from China. *Annales Botanici Fennici* **49**: 248–254.

Orel G, Wilson PG, Curry AS. 2014. Four new species and two new sections of *Camellia* (Theaceae) from Vietnam. *Novon* **23**: 307–318.

Orel G, Wilson PG, Curry AS, Luu HT. 2013. Two new species of *Polyspora* (Theaceae) from Vietnam and new combinations for some Asian species. *Willdenowia-Annals of the Botanic Garden and Botanical Museum Berlin-Dahlem* **43**: 301–308.

Prince LM. 1993. Theaceae. In: Committee FoNAE, ed. *Flora of North America*. New York: Oxford University Press, 322–323.

Prince LM. 2002. Circumscription and biogeographic patterns in the eastern north American–east Asian genus *Stewartia* (Theaceae: Stewartieae): insight from chloroplast and nuclear DNA sequence data. *Castanea* **67**: 290–301.

Prince LM, Parks CR. 2001. Phylogenetic relationships of Theaceae inferred from chloroplast DNA sequence data. *American Journal of Botany* **88**: 2309–2320.

Qiao X, Li Q, Yin H, *et al.* 2019. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biology* **20**: 38.

Quan C, Fu QY, Shi GL, *et al.* 2016. First Oligocene mummified plant Lagerstätte at the low latitudes of East Asia. *Science China Earth Sciences* **59**: 445–448.

Ran JH, Shen TT, Liu WJ, Wang PP, Wang XQ. 2015. Mitochondrial introgression and complex biogeographic history of the genus *Picea*. *Molecular Phylogenetics and Evolution* **93**: 63–76.

Ran JH, Shen TT, Wu H, Gong X, Wang XQ. 2018. Phylogeny and evolutionary history of Pinaceae updated by transcriptomic analysis. *Molecular Phylogenetics and Evolution* **129**: 106–116.

Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3**: 217–223.

Ronquist F, Teslenko M, van der Mark P, *et al.* 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* **61**: 539–542.

Rose JP, Kleist TJ, Lofstrand SD, Drew BT, Schonenberger J, Sytsma KJ. 2018. Phylogeny, historical biogeography, and diversification of angiosperm order Ericales suggest ancient Neotropical and East Asian connections. *Molecular Phylogenetics and Evolution* **122**: 59–79.

Sayyari E, Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. *Molecular Biology and Evolution* **33**: 1654–1668.

Schranz ME, Mohammadin S, Edger PP. 2012. Ancient whole genome duplications, novelty and diversification: the WGD radiation lag-time model. *Current Opinion in Plant Biology* **15**: 147–153.

Semerikova SA, Khrunyk YY, Lascoux M, Semerikov VL. 2018. From America to Eurasia: a multigenomes history of the genus *Abies*. *Molecular Phylogenetics and Evolution* **125**: 14–28.

Shi T, Huang HW, Barker MS. 2010. Ancient genome duplications during the evolution of kiwifruit (*Actinidia*) and related Ericales. *Annals of Botany* **106**: 497–504.

Shi XG, Fu QY, Jin JH, Quan C. 2017. Mummified Oligocene fruits of *Schima* (Theaceae) and their systematic and biogeographic implications. *Scientific Reports* **7**: 4009.

Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**: art31.

Smith SA, Moore MJ, Brown JW, Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology* **15**: 150.

Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. 2016. TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Research* **26**: 1134–1144.

Soltis DE, Gitzendanner MA, Stull G, *et al.* 2013. The potential of genomics in plant systematics. *Taxon* **62**: 886–898.

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

Stevens PF. 2001 onwards. *Angiosperm Phylogeny Website*. http://www.mobot.org/MOBOT/research/APweb/.

Stubbs RL, Folk RA, Xiang C-L, Chen S, Soltis DE, Cellinese N. 2020. A phylogenomic perspective on evolution and discordance in the Alpine-Arctic plant Clade *Micranthes* (Saxifragaceae). *Frontiers in Plant Science* **10**: 1773.

Takhtajan AL. 1997. *Diversity and the classification of flowering plants*. New York: Columbia University Press.

Tanaka T, Mizutani T, Shibata M, Tanikawa N, Parks CR. 2005. Cytogenetic studies on the origin of *Camellia* x *vernalis*. V. estimation of the seed parent of *C.* x *vernalis* that evolved about 400 years ago by cpDNA analysis. *Journal of the Japanese Society for Horticultural Science* **74**: 464–468.

Tang CQ. 2015. Evergreen broad-leaved forests. In: Tang CQ, ed. *The subtropical vegetation of Southwestern China: plant distribution, diversity and ecology*. Utrecht: Springer Netherlands, 49–112.

Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* **4**: 4.10.11–14.10.14.

Than C, Ruths D, Nakhleh L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* **9**: 322.

Tiffney BH. 1985. Perspectives on the origin of the floristic similarity between Eastern Asia and Eastern North-America. *Journal of the Arnold Arboretum* **66**: 73–94.

Tiffney BH, Manchester SR. 2001. The use of geological and paleontological evidence in evaluating plant phylogeographic hypotheses in the Northern Hemisphere tertiary. *International Journal of Plant Sciences* **162**: S3–S17.

Tsou CH. 1998. Early floral development of Camellioideae (Theaceae). *American Journal of Botany* **85**: 1531–1547.

Vekemans D, Proost S, Vanneste K, *et al.* 2012. Gamma paleohexaploidy in the stem lineage of core Eudicots: significance for MADS-box gene and species diversification. *Molecular Biology and Evolution* **29**: 3793–3806.

Wang YH, He H, Min TL, Zhou LH, Fritsch PW. 2006. The phylogenetic position of *Apterosperma* (Theaceae) based on morphological and karyotype characters. *Plant Systematics and Evolution* **260**: 39–52.

Wei CL, Yang H, Wang S, *et al.* 2018. Draft genome sequence of *Camellia sinensis* var. *sinensis* provides insights into the evolution of the tea genome and tea quality. *Proceedings of the National Academy of Sciences of the United States of America* **115**: E4151–E4158.

Wen DQ, Yu Y, Zhu JF, Nakhleh L, Posada D. 2018. Inferring phylogenetic networks using PhyloNet. *Systematic Biology* **67**: 735–740.

WFO. 2021. *World Flora Online*. Available at: http://www.worldfloraonline.org. Accessed 25 August 2021.

Wickett NJ, Mirarab S, Nguyen N, *et al.* 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences* **111**: E4859–E4868.

Wilde V, Frankenhauser H. 1998. The Middle Eocene plant taphocoenosis from Eckfeld (Eifel, Germany). *Review of Palaeobotany and Palynology* **101**: 7–28.

Xia EH, Tong W, Hou Y, *et al.* 2020. The reference genome of tea plant and resequencing of 81 diverse accessions provide insights into genome evolution and adaptation of tea plants. *Molecular Plant* **13**: 1013–1026.

Xia EH, Zhang HB, Sheng J, *et al.* 2017. The tea tree genome provides insights into tea flavor and independent evolution of caffeine biosynthesis. *Molecular Plant* **10**: 866–877.

Xiang YZ, Huang CH, Hu Y, *et al.* 2017. Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Molecular Biology and Evolution* **34**: 262–281.

Xu XD, Zheng W, Harris A, Wang W, Shao WZ, Wen J. 2018. Assessing the maternal origin in the polyploid complex of *Camellia reticulata* based on the chloroplast rpl16 intron sequences: implications for camellia cross breeding. *Molecular Breeding* **38**: 123.

Yan YJ, Davis CC, Dimitrov D, Wang ZH, Rahbek C, Borregaard MK. 2021. Phytogeographic history of the Tea family inferred through high-resolution phylogeny and fossils. *Systematic Biology* **70**: 1256–1271.

Yang JB, Yang SX, Li DZ, Lei LG, Ikeda T, Yoshno H. 2006. Phylogenetic relationships of Theaceae inferred from mitochondrial *matR* Gene sequence data. *Acta Botanica Yunnanica* **28**: 29–36.

Yang SX, Yang JB, Lei LG, Li DZ, Yoshino H, Ikeda T. 2004. Reassessing the relationships between *Gordonia* and *Polyspora* (Theaceae) based on the combined analyses of molecular data from the nuclear, plastid and mitochondrial genomes. *Plant Systematics and Evolution* **248**: 45–55.

Yang Y, Moore MJ, Brockington SF, *et al.* 2015. Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. *Molecular Biology and Evolution* **32**: 2001–2014.

Yang Y, Smith SA. 2013. Optimizing *de novo* assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics* **14**: 328.

Yang Y, Smith SA. 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Molecular Biology and Evolution* **31**: 3081–3092.

Ye CX. 1990. A discussion on relationship among the genera in Theoideae (Theaceae). *Acta Scientiarum Naturalium Universitatis Sunyatseni* **29**: 74–81.

Yu XQ, Drew BT, Yang JB, Gao LM, Li DZ. 2017*a*. Comparative chloroplast genomes of eleven *Schima* (Theaceae) species: insights into DNA barcoding and phylogeny. *PLoS One* **12**: e0178026.

Yu XQ, Gao LM, Soltis DE, *et al.* 2017*b*. Insights into the historical assembly of East Asian subtropical evergreen broadleaved forests revealed by the temporal history of the tea family. *New Phytologist* **215**: 1235–1248.

Yu XQ, Liu ED, Liu ZW, Xiao B, Ma JL, Yang SX. 2021. *Camellia luteocalpandria* (Theaceae), a new species and the first discovery of sect. *Calpandria* in China. *Phytotaxa* **489**: 223–228.

Yu XQ, Yang D, Guo C, Gao LM. 2018. Plant phylogenomics based on genome-partitioning strategies: progress and prospects. *Plant Diversity* **40**: 158–164.

Yu Y, Nakhleh L. 2015. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics* **16**: S10.

Zeng LP, Zhang N, Zhang Q, Endress PKH, Jie, Ma H. 2017. Resolution of deep eudicot phylogeny and their temporal diversification using nuclear genes from transcriptomic and genomic datasets. *New Phytologist* **214**: 1338–1354.

Zeng LP, Zhang Q, Sun RR, Kong HZ, Zhang N, Ma H. 2014. Resolution of deep Angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nature Communications* **5**: 4956.

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

Zhang CF, Zhang TK, Luebert F, *et al.* 2020. Asterid phylogenomics/phylotranscriptomics uncover morphological evolutionary histories and support phylogenetic placement for numerous whole genome duplications. *Molecular Biology and Evolution* **37**: 3188–3210.

Zhang W, Kan SL, Zhao H, Li ZY, Wang XQ. 2014. Molecular phylogeny of Tribe Theeae (Theaceae *s.s.*) and its implications for generic delimitation. *PLoS One* **9**: e98133.

Zhao L, Li X, Zhang N, *et al.* 2016. Phylogenomic analyses of large-scale nuclear genes provide new insights into the evolutionary relationships within the rosids. *Molecular Phylogenetics and Evolution* **105**: 166–176.

Zhao YY, Zhang R, Jiang KW, *et al.* 2021. Nuclear phylotranscriptomics/phylogenomics support numerous polyploidization events and hypotheses for the evolution of Rhizobial nitrogen-fixing symbiosis in Fabaceae. *Molecular Plant* **14**: 748–773.

Zwaenepoel A, Van de Peer Y. 2019. wgd-simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics* **35**: 2153–2155.