PHYLOGENY AND BIOGEOGRPHY OF THE TEA FAMILY

1 Title page

2 **Title: Phytogeographic history of the Tea family inferred through high-resolution phylogeny and**
3 **fossils**

4 Running title: Phylogeny and biogeography of the tea family

5 *Yujing Yan[1,2], *Charles C. Davis[2], Dimitar Dimitrov[1,3], Zhiheng Wang[4], Carsten Rahbek[5,1,4,6,7], Michael
6 Krabbe Borregaard[1]

7 *1. Center for Macroecology, Evolution and Climate, GLOBE Institute, University of Copenhagen,*
8 *Universitetsparken 15, 2100, Copenhagen, Denmark*

9 *2. Department of Organismic and Evolutionary Biology, Harvard University Herbaria, 22 Divinity Ave,*
10 *Cambridge, MA 02138, USA*

11 *3. Department of Natural History, University Museum of Bergen, University of Bergen, P.O. Box 7800, 5020*
12 *Bergen, Norway*

13 *4. Institute of Ecology, College of Urban and Environmental Sciences, Key Laboratory of Earth Surface*
14 *Processes of Ministry of Education, Peking University, Beijing 100871, China*

15 *5. Center for Global Mountain Biodiversity, GLOBE Institute, University of Copenhagen, Universitetsparken*
16 *15, 2100 Copenhagen, Denmark*

17 *6. Department of Life Sciences, Imperial College London, Silkwood Park campus, Ascot SL5 7PY, UK*

18 *7. Danish Institute for Advanced Study, University of Southern Denmark, Odense, Denmark.*
19
20 *** Corresponding authors:**

21 Yujing Yan, email: africarugu@gmail.com

22 Charles C. Davis, email: cdavis@oeb.harvard.edu

23

24

25

Abstract

The tea family (Theaceae) has a highly unusual amphi-Pacific disjunct distribution: most extant species in the family are restricted to subtropical evergreen broadleaf forests in East Asia, while a handful of species occur exclusively in the subtropical and tropical Americas. Here we used an approach that integrates the rich fossil evidence of this group with phylogenies in biogeographic analysis to study the processes behind this distribution pattern. We first combined genome-skimming sequencing with existing molecular data to build a robust species-level phylogeny for c.140 Theaceae species, resolving most important unclarified relationships. We then developed an empirical Bayesian method to incorporate distribution evidence from fossil specimens into historical biogeographic analyses and used this method to account for the spatiotemporal history of Theaceae fossils. We compared our method with an alternative Bayesian approach and show that it provides consistent results while significantly reduces computational demands which allows analyses of much larger datasets. Our analyses revealed a circumboreal distribution of the family from the early Cenozoic to the Miocene and inferred repeated expansions and retractions of the modelled distribution in the Northern Hemisphere, suggesting that the current Theaceae distribution could be the remnant of a larger continuous distribution associated with the boreotropical forest that has been hypothesized to occupy most of the northern latitudes in the early Cenozoic. These results contradict with studies that only considered current species distributions and showcase the necessity of integrating fossil and molecular data in phylogeny-based parametric biogeographic models to improve the reliability of inferred biogeographical events.

Key words

Genome skimming, plastid genome, Theaceae, phylogenomics, biogeography

The Pacific Ocean is the largest water basin on Earth and constitutes a formidable barrier for terrestrial species dispersal between tropical eastern Asia and the neotropical region in the Americas, yet several plants and animals are found in both regions. Such disjunct distributions, where a

monophyletic lineage occurs on both the eastern and western edges of the Pacific basin (i.e., temperate to tropical Asia on the one side and southeastern North America and South America on the other), with no occurrences in between are referred to as 'amphi-Pacific'. Amphi-Pacific distribution is exhibited by more than 100 genera and a few higher taxa within angiosperms (Steenis 1962; Thorne 1972). In spite of this relatively high number of occurrences, the origin of this distribution remains controversial. Unravelling the causes of the amphi-Pacific distribution will have key implications for our understanding of the footprint of deep-time historical processes on present plant biogeographical patterns.

Amphi-Pacific distributions have been hypothesized to be relicts of wider circumboreal distributions associated with a continuous belt of 'boreotropical' evergreen forest, which is thought to have extended at middle to northern latitudes through Eurasia and America during the early Cenozoic. This forest was most likely continuous via a North Atlantic Land Bridge and/or the Bering Land Bridge, supported by evidence from both plants and animals (Tiffney 1985b, 1985a; Lavin and Luckow 1993; Sanmartín et al. 2001; Davis et al. 2002; Condamine et al. 2013). According to paleobotanical evidence, it was replaced by mixed mesophytic forest around early Oligocene and later boreal forest in late Miocene following climate cooling, resulting in the extinction and/or southward migration of boreotropical thermophilic taxa at high latitude regions (Meseguer et al. 2015, 2018). Evidence of adaptation to more temperate climate was also found (Meseguer et al. 2018). Recent analyses have inferred such a boreotropical forest origin for several plant groups with amphi-Pacific distributions, based on dated phylogenies and ancestral range reconstructions (e.g., Antonelli et al. 2009; Li et al. 2011a; Li and Wen 2013; Fritsch et al. 2015; Xiang et al. 2016), though biogeographical analyses are inherently problematic for lineages with no current members in intervening regions. The most important alternative hypothesis proposes an origin of the group in either Eastern Asia or North America followed by one or several long-distance dispersal events,

either via the Bering Land Bridge or by sea currents across the Pacific Ocean (Wen et al. 2010; Christenhusz and Chase 2013; Wu et al. 2018).

One of the best-known plant lineages exhibiting an amphi-Pacific distribution is the tea family (Theaceae). Placed in Ericales, the family contains three tribes, ca. nine genera (Prince 2007), and 368 accepted species according to The Plant List (TPL, 2013), comprising shrubs and trees that are mostly thermophilic species and inhabit broadleaved-evergreen forests. Within the family, all three tribes have a disjunct amphi-Pacific distribution. *Camellia*, *Schima*, *Pyrenaria*, *Polyspora*, *Apterosperma*, and most members of *Stewartia* (including *Hartia)* and *Gordonia* are restricted to subtropical and tropical Asia, whereas ca. 20 species belonging to several morphologically distinct groups, including *Stewartia*, *Gordonia*, *Laplacea* and *Franklinia,* are restricted to the southeastern North America and the Neotropics, with no occurrences in-between.

Several attempts have been made to understand the evolutionary history of the family using biogeographical reconstructions. A recent study based on a relatively species-sparse phylogeny of Ericales inferred the stem of the family to be of Indo-Malaysian origin, with a possible expansion into the Nearctic at ~63 Ma (Rose et al. 2018), supporting an earlier conjecture by Li et al. (2013). The crown groups for all three tribes (Stewartieae, Gordonieae and Theeae) were all estimated to have originated during the late Oligocene to mid-Miocene (Yu et al. 2017; Lin et al. 2019), with uncertain biogeographical origin. The most recent common ancestors of both Stewartieae (*Stewartia s.l.*) and Gordonieae may have originated in North America, with subsequent dispersal events into East Asia involving multiple species during the Miocene (Lin et al. 2019). In particular, *Gordonia s.l.* may have species on both sides of the Pacific. The relationships within this genus have been contested by several authors, and even its status as a monophyletic clade is in question (Prince and Parks 2001; Yang et al. 2004; Gunathilake et al. 2015). The Neotropical distribution of *Laplacea* in Theeae, on the other hand, has been argued to result from a single long-distance dispersal event from

Asia to South America (Li et al., 2013), though the position of key taxa involved in the disjunct distribution is uncertain.

Poorly resolved relationships within the family and the sparse sampling of relevant species may lead to underestimating rates of cladogenesis and potentially erroneous conclusions in ancestral area reconstruction (Meseguer et al. 2015). Other potentially promising approach to improve biogeographic reconstructions and to decrease their level of uncertainty is to draw on fossils that directly reveal past occurrences. Several recent studies have experimented with including fossil taxa with reliable phylogenetic positions and spatial location data into parametric biogeographic models when reconstructing range dynamics. In some cases, including fossil information has led to substantial changes in the inferred biogeographic scenarios (Mao et al. 2012; Nauheimer et al. 2012; Wood et al. 2013). Thus, here we have made specific effort to improve not only the sampling of extant species, but also of fossil species.

However, accurately placing fossils onto the phylogeny usually takes a total-evidence approach which requires: 1) fossils with very informative morphological characters, 2) a comparative and well-sampled morphological character matrix for extant species, and 3) preferably a congruent evolutionary history of sampled morphological characters and molecular evidence. Such data are not always feasible for most clades including our target group and analyses using this approach have so far been restricted to a few small clades with well-preserved fossils (Meseguer and Condamine 2017a). Theaceae is known from many macrofossils dating as far back as the late Cretaceous, occurring through the Cenozoic across temperate regions in the Northern Hemisphere (Fig. 1; Grote and Dilcher 1989, 1992; Bozukov and Palamarev 1995). Interestingly, these fossils from mid to high latitude may reflect a wide and northern distribution of the group and subsequent extinctions as many of them occur far north of the distribution of extant lineages (Sanmartín and Meseguer 2016a). This hypothesis has only been proposed by paleobotanists for the genera *Gordonia* and *Schima* based on

the visual appearance of fossils (Grote and Dilcher 1992; Shi et al. 2017) but has not been tested quantitatively.

(Figure 1.)

In this study, we generated a robust time-calibrated species-level phylogeny of Theaceae with wide taxonomic and genomic coverage, using *de-novo* sequenced and previously published chloroplast genomes and nuclear ribosomal DNA (nrDNA) sequences for 146 species in nine genera (in the broad sense) across the world. Sequencing focused particularly on problematic and undersampled groups such as *Gordonia s.l.* and *Camellia*. We developed a heuristic method to incorporate fossils into the phylogeny based on their taxonomic placement and estimated ages even when they cannot be confidently placed on the phylogeny using morphological characters. We then use an empirical-Bayesian approach to infer biogeographical history that accounts for the phylogenetic uncertainties and evaluates the effect of incorporating information from extinct taxa. To resolve the enigmatic amphi-Pacific distribution of Theaceae, we (1) provide a new reference phylogeny for the tea family for the biogeographic analysis, (2) clarify the placement of the problematic group *Gordonia* and its relationship to *Laplacea* and *Polyspora,* (3) evaluate the effect of incorporating fossil information for the reconstructed biogeographic history.

MATERIALS AND METHODS

Our analytical approach involved building an improved phylogeny of the Theaceae by sequencing and including in the phylogenetic analyses species that have been difficult to place. For phylogenetic inference, we chose chloroplast genomes (pt, plastome) and nuclear ribosomal high-copy regions (nrDNA: the 18SrRNA-ITS1-5.8SrRNA-ITS2-26SrRNA region), because these markers have resolved relationships within the family well in previous studies and data for c.60 of species is already available (Yu et al. 2017b). We assembled target regions for 83 species *de novo* from herbarium specimens following genome skimming methods by Marinho et al. (2019). We then combined these data with existing plastid genomes and nrDNA data downloaded from GenBank and

inferred a new dated phylogeny aiming to cover the entire family. We developed a novel method that allow the incorporation of information on past distributions based on the fossil records in biogeographic reconstruction.

Taxonomic Sampling, DNA Sequencing, and Data Processing

We sequenced *de novo* 83 herbarium specimens using the genome skimming method (Alamoudi et al. 2014; Dodsworth 2015; Zeng et al. 2018; Marinho et al. 2019). We focused on specimens from problematic or poorly covered taxa or regions, in particular *Gordonia s.l.* and *Camellia* (a vouchered specimen list is presented in Table S2). For each species, we selected the most recently collected specimen from within the native range of the species. Taxonomic assignments of specimens were based on the most recent name identified on the specimen sheets. Combined with the GenBank dataset, the resulting species coverage for all genera except *Camellia* was higher than 80% and the coverage within *Camellia* was c. 31% (according to TPL 2013) or c. 64% (according to Flora of China 2007).

Approximately 15 mg of leaf tissue was used for each DNA extraction. Both the Promega Maxwell kit (customized for herbarium specimens following the manufacturer's instructions) and a modified CTAB protocol (Doyle and Doyle 1987) were used to extract DNA from specimens. All extractions were quantified using a Qubit® 3.0 Fluorometer (Life Technologies, Carlsbad, California, USA) and quality checked using Nanodrop ND-1000 Spectrophotometer (Thermo Fisher Scientific). Samples collected at different times in the past were selected to visually assess the integrity with an Agilent Technologies 4200 TapeStation System using the Genomic DNA ScreenTape (Agilent Technologies, Santa Clara, California, USA).

Dual-indexed libraries were prepared with 2-48ng input genomic DNA using the Kapa DNA Hyper Plus Library Prep Kit (Kapa Biosystems) at 1/4 the recommended volume and size selected for 250-700bp. The final libraries were pooled in equimolar rations. The samples were sequenced either on Illumina NextSeq 500 mid output platform with 2*150-bp paired-end reads or Illumina

HiSeq 2500 platform with 2*125-bp paired-end reads at the Bauer Core Facility of the Harvard University (https://www.rc.fas.harvard.edu/odyssey-3-the-next-generation/).

Adapters were trimmed, low quality reads were removed, and base correction was performed using the default setting in fastp (Chen et al. 2018). Samples with > 100,000 filtered reads were processed in Geneious R11 (https://www.geneious.com) with medium-low sensitivity and up to ten iterations to assemble plastome and nuclear DNA sequences. For plastome assembly, we used the complete plastome of *Camellia taliensis* (NC_ 022264.1) as reference sequence. For nrDNA, nrDNA of *Camellia elongata* (MF171090.1) was used as reference sequence. For each region, two consensus contigs were generated using a 75% masking threshold and highest chromatogram quality. To avoid sequencing errors and potential contamination, bases with low coverage (below 6X for the plastome and 11X for nrDNA) were masked with Ns (Ripma et al. 2014). Samples with < 40% high-quality bases, > 60% ambiguities and an average sequencing depth < 5 were excluded in the following analysis. Raw read data generated in the study are archived in the NCBI Sequence Repository Archive.

Building the molecular dataset from Genbank data

We downloaded all available plastomes and nrDNA of Theaceae from GenBank to supplement our sequencing dataset to achieve as complete a sampling of the family as possible at the species level. To ensure the reliability of our data, we only included plastome sequences cited in publications. As there is no widely accepted classification for Theaceae worldwide, our taxonomy followed the Plant List (2013) and excluded taxa that were listed as unresolved. The cleaned GenBank dataset included 69 complete plastomes and 41 nrDNA sequences for 70 species (Table S1).

Additionally, we downloaded chloroplast genomes and nrDNA sequences in the RefSeq dataset for 19 species of various families within Ericales as outgroups. The selection of these

outgroups mostly followed Yu et al. (2017b) with additional species sampled in Primulaceae and Pentaphylaceae (Table S1).

Dataset composition and Alignment

The final dataset was assembled by combining the genome skimming data with those downloaded from GenBank (Table S1 plus Table S2). The dataset comprised 146 species.

We constructed two subsets of these data to assess the congruence of phylogenies reconstructed from different genomic compartments, i.e., plastid coding regions (CDS) and nrDNA. To assemble the CDS dataset, we extracted the coding regions of all plastomes using BLAST+ (Camacho et al. 2009) against four annotated reference genomes, i.e., *Camellia taliensis* (NC_022264.1), *Schima brevipedicellata* (NC_035537.1), *Euryodendron excelsum* (NC_039178.1) and *Ardisia polysticta* (NC_021121.1) to eliminate possible annotation errors.

We used a partition strategy for the alignment to reduce the impact of sequencing and assembly errors. Sequences were roughly aligned using MAFFT v.7 (Katoh et al. 2002), then partitioned based on gene position as extracted from the annotations of 49 plastomes in RefSeq database (sequence accession started with "NC_" in Table S1, including four outgroup species) and the annotations of 54 nrDNA sequences (sequence accession started with "MF" in Table S1, including 12 outgroup species). This resulted in 105 partitions for the chloroplast genome and five partitions for the nrDNA (Table S3). Each partition was aligned separately, and the resulting alignments were used to infer phylogenies with FastFree (Price et al. 2010). For each partition, we removed sequences with a threshold of 0.05 in TreeShrink (Mai and Mirarab 2018) to reduce long branch attraction artifacts (LBA). The filtered partitions were then concatenated and trimmed to exclude indel-rich positions using the "auto" setting in TrimAl (Capella-Gutiérrez et al. 2009).

Phylogenetic Analysis and Divergence Time Estimation

We concatenated the plastome and nrDNA datasets to reconstruct the full phylogeny using partitioned maximum likelihood method, which allows each data partition (plastome and nrDNA) to

have different substitution models (comparable to the approach taken by Yu et al. 2017b). The best nucleotide substitution model for each partition was determined with PartitionFinder2 (Lanfear et al. 2016), based on the corrected Akaike Information Criterion (cAIC). The partitioned maximum likelihood analysis was accomplished with the MPI version of RAxML-ng (Kozlov et al. 2019). Clade support was estimated using non-parametric bootstrap analyses with 1000 replicates and bootstrap values were mapped to the best-scoring tree. In this process, we observed several unstable taxa (rogue taxa) that affected clade support negatively (Wilkinson, 1996). We qualified the influence of each species on the stability of the phylogeny using RogueNaRok (Aberer et al. 2013) (Table S4), pruned the rogue taxa and reran the above analysis using the reduced dataset.

To evaluate the stability of the topology to dataset selection and phylogenetic methodology, we also built phylogenies for the CDS and the nrDNA datasets separately using both maximum likelihood and a Bayesian inference framework. For the maximum likelihood analysis, we used the same parameter settings as mentioned previously, but used 200 pseudo-replicates for a bootstrap analysis of branch support. For the Bayesian analysis, we used MrBayes v3.2.6 (Huelsenbeck and Ronquist 2001). Two independent runs were conducted with four Markov chains (one cold and three heated) for $> 3*10^7$ generations, and a sampling frequency of one tree every 300 generations. Convergence was assumed when the average standard deviation of split frequencies was < 0.01. We summarized the posterior on the best maximum likelihood tree, to facilitate comparison between the results of the two methods.

Due to the large size of the phylogeny, we used an efficient penalized likelihood method implemented in treePL (Smith and O'Meara 2012) to estimate divergence times within the trees generated by RAxML-ng. We selected three fossil calibration points for the outgroups, following Yu (2017b), and four fossil calibration points for the ingroup species (Fig. 2, TableS5, Li et al. 2013; Yu et al. 2017b). We conservatively applied all calibration constraints to the stems of the groups where respective fossils were placed. Penalized likelihood uses a smoothing parameter to accommodate rate

heterogeneity. We used Random Subsample and Replicate Cross-Validation (RSRCV) to determine the appropriate smoothing parameter for our dataset. To assess uncertainty in age estimates we estimated confidence intervals on inferred ages by dating all 1000 ML bootstrap trees for the concatenated dataset. Results from the dating of the bootstrapped trees were then summarized and resulting confidence intervals were visualized on the best tree using TreeAnnotator (part of the BEAST2 package, Bouckaert et al. 2014).

All phylogenetic analyses were implemented on the CIPRES Science Gateway (Miller et al. 2010) or the Harvard cluster Odyssey (https://www.rc.fas.harvard.edu/odyssey-3-the-next-generation/).

Selection and Incorporation of Extinct Taxa

To evaluate the impact of fossil occurrences on biogeographic inference, we developed a method to directly incorporate the fossil distribution information in phylogeny-based parametric biogeographic models by using phylogenies with fossils assigned iteratively based on their taxonomic placement. We compiled a comprehensive dataset of all fossils of Theaceae from the Paleobiology Database (http://paleodb.org), Japan Paleobiology Database (http://jpaleodb.org/) and published literature (see Fig. 1 for the distribution of all fossil records). Online databases do not provide confirmation of identification and might include misplaced fossils. To limit the uncertainties, we first filtered the dataset to unique and most reliable records by applying the following criteria: 1) the fossil has a locality at city or provincial level, 2) the fossil is dated to a geological epoch, 3) the fossil is assigned to a modern genus, 4) the fossil is associated with a publication later than 1990, 5) the fossil is a macrofossil, and 6) there is a detailed description with comparison to modern analogues. After applying these filters, we ended with 22 unique fossil records with species names (detailed information of the fossils and arguments that support the taxonomy are presented in Table S7).

It is often difficult to place woody plant fossils accurately on a phylogeny due to missing information on some of the morphological characters and the lack of comparative morphological character matrices for extant species. To address this issue, we implemented a heuristic approach that assigned fossils randomly to a node within the most recent clade a fossil could be assigned to with certainty (usually a genus). The fossils were added as extinct lineages terminating within the confidence interval of the deposition time of the fossil, defined as the interval from the maximum to the minimum possible age of the formation in which the fossil was found (according to the latest geological time scale, Walker et al. 2018).

To place each fossil on the phylogeny, we first identified the narrowest clade that it could be assigned to with certainty, following the taxonomic placement of the most recent publication describing the fossil (see Table S7 for more details). We then selected a branch within this clade and added the fossil specimen as a side branch to it. The branch was selected among a candidate set of all branches in the clade that existed prior to the minimum age of the fossil (based on the time interval given in the literature for the youngest specimen of that species). The new side branch was added to the selected branch at a randomly chosen time, drawn from a uniform distribution along the branch (but no later than the minimum age of the fossil). The terminal age of the fossil was then drawn from a uniform distribution between the newly created node and the fossil's minimum age. Note that this procedure may cause fossils to branch off from the stem above their assigned genus, effectively increasing the crown age of the genus for subsequent fossils. For this reason, we added the fossils in order from the oldest to the youngest, making it possible to assign younger fossils to the stem nodes created by older fossils.

One potential complication with this approach is that several fossils of morphologically highly similar species, in many cases found together, may be placed in different locations by the algorithm, though the most parsimonious explanation would be that they belonged to the same radiation. Dispersing these species over the phylogeny may lead to unrealistic ancestral state

297 reconstruction and overestimation of dispersal events. To address this issue, we subsampled the
298 fossil dataset so that only one species per geological age per area was retained. We excluded fossils
299 found in the Sino-Japanese region, which is well represented by present taxa in space. Applying
300 these additional criteria resulted in a final set of 10 fossils (Table S7, Fig. S3).

301 *Biogeographic Analysis with and without Fossil Taxa*

302 Biogeographic analyses require the a priori definition of distinct regions. We defined large
303 regions based on the phyloregions defined by Holt et al. (2013), merging their regions into five
304 areas: (1) Eurasian (E), (2) Nearctic (N), (3) Sino-Japanese (S), (4) Panamanian (P), (5) Papua-
305 Melanesian (M). Africa and Australia were excluded from the analysis, as there are no reliable
306 fossils or current occurrences of Theaceae in these regions. Moreover, the age of the crown node of
307 Theaceae was estimated to be around 60 Ma (with upper 95% highest posterior density boundary of
308 74.7 Ma) (Yu et al. 2017b) and is much younger than that expected for a Gondwanan origin (>80Ma)
309 (Beaulieu et al. 2013). Eurasian here represents the high latitude region of the Eurasian landmass,
310 which used to be connected to the Nearctic through the North Atlantic land bridge or the Bering land
311 bridge in historical times. The boundaries of this region were set based on the phylogenetic
312 regionalization of Holt et al. (2013). Each species was assigned to one or more of these regions in
313 ArcGIS 10.5 (Esri Inc., 2016) based on species' occurrence data collected from multiple data sources
314 (see Supplementary Appendix1), as well as from specimen tags from the species sequenced for our
315 phylogenetic analysis. We limited the occurrences to specimen records, and cleaned the occurrences
316 using CoordinateCleaner (Zizka et al. 2019).

317 The biogeographical history of the family was inferred using a likelihood-based dispersal-
318 extinction-cladogenesis (DEC) model and a DEC model with founder events parameter (DEC+j)
319 implemented in R package BioGeoBEARS (Ree and Smith 2008; Matzke 2013). Both methods were
320 used because DEC+J often artificially results in higher likelihood values than DEC, so the
321 likelihoods themselves are not adequate for model selection in that case (Ree and Sanmartin 2018).

The BioGeoBEARS DEC and DEC+j models allow inclusion of fossil tips in the phylogeny and inference of extinction directly from biogeographic data, which is an advantage over cladistic and event-based methods (Matzke 2013; Sanmartín and Meseguer 2016b). The models were run with some key assumptions that cannot be determined a priori from our dataset and may affect the final inference. We carried out a thorough sensitivity analyses and evaluated a larger set of feasible parameter combinations. The set of parameters that made the largest difference in terms of conclusions are presented here in the main text, the rest in the supplementary materials. These parameters are whether or not to allow founder event dispersal (the DEC vs DEC+j model), the maximum number of occupied regions (set to two or three), and whether dispersal probabilities between Nearctic and Eurasian should vary over time to reflect the disappearance of the boreotropical forest corridor (yes or no). (Mao et al. 2012; Fritsch et al. 2015; Meseguer et al. 2015; Rose et al. 2018).

Basic dispersal probabilities among regions assumed that species could move between Nearctic and Eurasia regions freely, but only disperse to southern regions from their adjacent region (M0). When evaluating the effect of allowing inter-region dispersal probabilities to vary, we applied a time-stratified matrix with three time slices (70-35Ma, 35-10Ma, 10-0Ma), where dispersal probabilities reflected the connectivity of Northern Hemisphere regions for the tea family following the vegetation reconstruction in Meseguer et al. (2015) (M1, Table S6). We set three probability categories: 0.01 for well-separated areas, 0.5 for moderately separated areas, 1.0 for well-connected areas. We used categories rather than an actual distance matrix in the analysis because distances among regions change continuously and are difficult to quantify.

In order to accommodate phylogenetic uncertainties, we used an empirical Bayesian approach inspired by Nylander et al. (2008), in which the biogeographic modeling procedure was repeated on 300 phylogenies randomly sampled from the 1000 dated bootstrap trees and the average marginal regional probabilities ware summarized on the best RAxML-ng tree in BioGeoBEARS

followed Smith (2009) and Cai et al. (2016). To evaluate the effect of including fossil distributions in biogeographic analysis, we assigned the two sets of fossils to each of the 300 trees using the procedure described above, ran the models with same settings, and summarized the results on the best RAxML-ng tree. We also summarized the main parameters of the different models and evaluated model performance using AICc and likelihood-ratio tests (lrt).

Averaging marginal regional probabilities from the sampled trees to the best tree only considers results of identical clades and does not necessarily reflect the statistical distribution of different evolutionary histories. To evaluate the ancestral ranges at nodes of major clades and possible historical events along branches of the sampled trees, we performed the newly developed biogeographical stochastic mappings procedure (BSM) (Dupin et al. 2017) for each tree in BioGeoBEARS simulating the biogeographic history based on the corresponding biogeographical likelihood model. The BSM realizations were summarized after 50 successful BSMs in 10000 tries. We finally calculated the proportions of the most likely ancestral ranges for major clades and summarized the number of different events across the 300 sampled trees. See Figure S1 for an illustration of the workflow of our sequential empirical-Bayesian analysis.

All the biogeographical analysis was performed in R 3.5.3 (R Core Team, 2019). Phylogenies were plotted using the R package ggtree (Yu et al. 2017a).

RESULTS

Phylogenomics of Theaceae

We succeeded in assembling at least 60% of the plastome for 59 samples (median sequencing depths at 29X) and the entire ribosomal sequences cluster for 72 samples (median sequencing depths at 278X) from the genome skimming data (Table S2). Two New World species (*Gordonia haematoxylon* and *Laplacea wrightii var. moaensis*) were represented by two samples each. As both samples grouped together with high support values for both species, only one of each was kept for the combined plastome-nrDNA dataset. The proportion of parsimony-informative characters for

plastid coding regions (CDs), nrDNA and concatenated plastome-nrDNA datasets were 14%, 14% and 20%, respectively, with alignment lengths of 68867bp, 6142bp, and 135216bp. In the phylogeny based on the combined dataset, nine species were removed after being identified as rogue taxa (Table S4), for a total of 147 species in the final dataset (128 ingroup species and 19 outgroup species). (Figure 2.)

All datasets recovered the genera *Pyrenaria*, *Stewartia* and *Schima* as monophyletic, in accordance with previous molecular phylogenetic studies. Conversely, paraphyly of *Gordonia s.l.* and *Laplacea* was suggested with high support values across all the datasets. In the phylogeny based on the combined dataset, only *Gordonia brandegeei (synonym: Laplacea grandis)* was retained in the same clade as the type species for *Gordonia*, *Gordonia lasianus*. This reduced *Gordonia* clade was strongly supported as the sister clade to *Franklinia alatamaha* + *Schima* with high bootstrap support (BS 100%). The following relationships were also supported by 100% BS: all the central American species, including *Laplacea fruticosa*, *Laplacea angustifolia, Laplacea portoricensis, Laplacea wrightii var. moaensis,* and *Gordonia haematoxylon* (*Laplacea haematoxylon*), formed a monophyletic clade, which was sister to *Apterosperma oblata*. *Gordonia yunnanensis* was placed in *Pyrenaria*. *Gordonia szechuanensis* was placed in a previously recognized *Polyspora* clade while all the Southeast Asian *Gordonia* species and *Gordonia balansae* formed a monophyletic clade and was sister to *Polyspora*. (Fig. 2, Fig. S2).

The topologies were consistent across different phylogenetic reconstructions, and both maximum likelihood and Bayesian analysis inferred similar topologies for the CDs and nrDNA datasets (Fig. S2-S3). The combined dataset had the highest overall support and resolution. The highly supported nodes (BS> 70% and PP > 0.95) of the nrDNA phylogeny were mostly consistent with the CDs phylogeny and the combined plastome-nrDNA phylogeny, though the overall nodal support was low (average 51% BS). Despite high congruence among datasets of inferred relationships within the three tribes (Stewartieae, Gordonieae and Theeae), the relationships among

397 them differed between datasets. This incongruence was also reflected by moderate support for
398 Theeae and Gordonieae as sister taxa, with Stewartieae as a sister group to that clade in the
399 combined dataset. Our biogeographical analyses used an empirical Bayesian method to explicitly
400 account for such uncertainties in the phylogenetic reconstruction.

401 The backbone relationships between some clades remained poorly resolved (Fig. S2-S3).
402 Notably, *Camellia gracilipes* grouped with the *Polyspora*+Asia *Gordonia* clade with high support in
403 the plastid dataset but was placed within *Camellia* in the analysis based on nrDNA data with very
404 low support.

405 *Divergence Time Estimation*

406 For the treePL analyses we set the maximum root depth to 125 Ma. Changing this setting did
407 not give qualitatively different results for the internal node ages (Fig. 2 and Table S8). The stem age
408 of Theaceae was estimated to be 92.1 Ma (95% confidence interval (CI), 84.4-99.8) and the crown
409 age 66.4 Ma (95% CI: 60.0-73.4). The crown age of Stewartieae (including the North American
410 *Stewartia malacodendron*) was 22.69 Ma (95% CI: 15.1-28.7). The stem and crown ages of New
411 World *Gordonia* were 25.6 Ma (95% CI: 24.1-26.4) and 20.8 (95% CI: 12.6-22.8). The stem and
412 crown ages of *Laplacea*, another New World lineage, were 24.1 Ma (95% CI: 21.2-29.3) and 12.9
413 (95% CI: 8.7-16.1). The estimated ages were all within the 95% highest posterior density intervals of
414 Yu et al. (2017b) and Rao et al. (2018).

415 *Biogeographical Analysis with BioGeoBEARS*

416 The model allowing jump dispersal (DEC+J) had higher likelihood values than the
417 vicariance-only model (DEC) for all three phylogenetic datasets and across parameter settings (Table
418 1 and Table S9). However, there are arguments against model selection using likelihood-based
419 methods in DEC-based models (Ree and Sanmartín 2018). For plants, jump dispersal is a strong
420 assumption to take in biogeographic analysis, in that it assumes an ability to jump between remote

geographical regions multiple times, and as such a vicariance-only model may be considered more parsimonious. Consequently, we present the results from both models.

(Table 1)

The model parameters were congruent for all datasets under both M0 and M1 dispersal scenarios, and the reconstructed range dynamics were almost identical. Therefore, we focus on the results of the M0 scenario. Several observations about parameter estimates were consistent with the results of Ree and Sanmartín (2018). The dispersal rate of the DEC model was estimated to be higher than the DEC+j model for all datasets. Incorporating fossils into phylogenies increased the dispersal rate and extinction rate for DEC models, but did not influence these rates in DEC+j models. The extinction rate of DEC+j models stayed zero regardless of the number of fossils added, whereas the rate of founder events increased significantly (average 0.01 to 0.045 and 0.055 when max range was set to two) (Table 1). Similar patterns were found when the maximum number of co-occupied regions was increased from two to three (Table S9).

Biogeographic Reconstruction without Fossil Taxa

All models that did not include fossil taxa reconstructed similar ancestral range dynamics regardless of dispersal settings, with a discontinuous distribution of the crown node (Fig. 3a and Fig. S5-Fig. S7). These findings indicate that the dispersal matrix had a very limited impact on the ancestral state reconstruction, as also found by Chacón and Renner (2014). Very few dispersal and vicariance events were recovered and most were inferred to have occurred around the transition between boreotropical forest to mixed mesophytic forest (Fig. 4a). Most dispersal events were from Sino-Japanese regions (Fig. 3c). Analyses reconstructed the crown node of Theaceae as Nearctic+Sino-Japanese (NS) for 60-98% of the sampled phylogenies (Fig. 3a and Table S10).

The ancestral areas estimated for the crown of Stewartieae were NS as well for the majority of the sampled phylogenies, while the most recent common ancestor of Theeae and Gordonieae was Sino-Japanese region only. Within Stewartieae, *Stewartia malacodendron* was reconstructed as

446 diverging in the Nearctic in the early-Miocene, with gene flow between New World and Old World
447 ceasing around the mid-Miocene, after which the lineage experienced a vicariance event that left
448 *Stewartia ovata* in the Nearctic while its sister clade diversified in the Sino-Japanese region (Fig. 3a).
449 The most recent common ancestor of the Theeae and Gordonieae tribes was reconstructed as
450 expanding from the Sino-Japanese region into the Nearctic, again placing the crown of tribe
451 Gordonieae with an NS distribution. During the late Oligocene and mid-Miocene, one descendant of
452 Gordonieae expanded south to the Panamanian region (Fig. 3a). The distribution of the crown node
453 of Theeae differed between the DEC+J model and the DEC model. With jump dispersal, the crown
454 node was placed in the Sino-Japanese region, indicating that the Neotropical distribution of the genus
455 *Laplacea* must have been the result of a later long-distance dispersal event across the Pacific Ocean
456 (Fig. S5). A joint occupation of the Sino-Japanese+Panamaian (PS) region was inferred under the
457 DEC model (Fig. 3a).
458 (Figure 3.)
459 *Biogeographic Reconstruction including Fossil Taxa*
460 Adding fossil taxa radically altered the reconstructed geographical distributions for the deep
461 nodes of the family. Analysis including 10 fossil taxa placed the family in Eurasia from the early
462 Cenozoic to the mid-Miocene (Fig. 3b). Under the M0 dispersal scenario, the DEC model inferred
463 the ancestral range of crown Theaceae to be either Eurasian+Nearctic (EN, ~50% of the phylogenies)
464 or Eurasian (~33% of the phylogenies). The reconstruction involved multiple vicariance and
465 dispersal events between Eurasian and the Nearctic and dispersal events between Eurasian and the
466 Sino-Japanese region. Eurasia was inferred as a major source region other than Sino-Japanese region
467 (Fig. 3d). The number of events recovered was higher than reconstruction based only on extant taxa
468 and the highest number of events happened within the period of mixed mesophytic forest (Fig. 4b).
469 Similar results were obtained under the M1 scenario (Fig. S6b).

The crown-group of Stewartieae was inferred to be originally widespread in the EN regions (supported by more than 66% of the phylogenies), followed by expansion into the Sino-Japanese and subsequent extinction in the Nearctic and Eurasia from the mid-Oligocene to mid-Miocene. The crown-group Gordonieae exhibited a similar pattern, with a slightly more complex history for the Theeae, with an earlier Eurasian extinction around the early Oligocene with a possible temporary recolonization in the Early Miocene. The ancestral occurrence of extant crown-group Theeae was the same as in the reconstruction without fossils. When including jump dispersal events, ~50% of the phylogenies favored Eurasia as the ancestral range, while 33% favored EN. Multiple jumping events between Eurasia and Nearctic, as well as between Eurasia and the Sino-Japanese region replaced the inferred vicariance in the DEC model. The crown-group of Stewartieae was estimated to have occupied Eurasia in more than 60% of the phylogenies, while that of Gordonieae was estimated to have occupied the Nearctic for most phylogenies (Fig. 3b).

Analysis on the dataset including 22 fossils also favored NS for the crown-group Theaceae in 50-60% of the sampled phylogenies, with 15-26% of the samples favoring a Eurasian origin (under the assumption that no clade can occupy more than two regions at any given time; (Table S10). The crown group of Stewartieae was assigned to ES with high probability. The Sino-Japanese region was inferred as a major source region, with more than four dispersal events inferred from this region to the Euraisan, on average (Fig. S8-9). However, including many fossils with unresolved phylogenetic position into analyses may lead to some counter-intuitive reconstructions and increase the sensitive to dispersal constraints. In the M1 dispersal scenario, only 55% of the samples succeeded in 50 realizations in 10000 tries of BSM under the DEC model.

Relaxing the assumptions to allow occupancy of three regions simultaneously broadened the inferred occupancy of the family crown group somewhat, to either encompass the Eurasia-Nearctic-Sino-Japanese or the Eurasia-Panamanian-Sino-Japanese regions (Table S11).

(Figure 4.)

DISCUSSION

We provide a combined plastome and nrDNA phylogeny for Theaceae with high taxon coverage. Our phylogeny greatly improves the understanding of Theaceae relationships, especially of *Gordonia s.l.* and *Laplacea*. Our results demonstrate that biogeographical analyses based solely on extant species distributions substantially misrepresented the past distributional history for the clade. This was remedied by adding a small number of fossil specimens and despite the high uncertainty of their phylogenetic placement. Biogeographical analysis accounting for known fossil distributions revealed a boreotropical origin for the family, with the amphi-Pacific disjunct distribution arising from multiple colonization events from north to south and subsequent extinction in the intervening areas of Eurasia. The biotic exchange between the Old and the New World started in the early Eocene and ended in the Miocene. The study highlights the vital importance of the targeted sequencing of problematic taxa and inclusion of fossil data for robust biogeographical analyses.

Phylogenetic Relationships of Theaceae and Gordonia s.l.

Both plastome and nrDNA dataset support paraphyly of the problematic genus *Gordonia s.l.*, indicating that it should be divided into three clades as previously proposed by Prince and Parks (2001) and Yang et al. (2004). The *Gordonia* clade only includes the type species, *Gordonia lasianus*, and *Gordonia brandegeei* (synonym: *Laplacea grandis*). This forms a purely New World clade, distributed from the Southeastern U.S. south to Columbia. The Caribbean and South American species, which include the genus *Laplacea* and the species *Gordonia haematoxylon* (synonym: *Laplacea haematoxylon*), formed a robust clade within the Theeae tribe. All the Asian species of *Gordonia* were placed in the genus *Polyspora* of tribe Theeae, putatively placed as the sister group to *Camellia,* corroborating some previous studies (Yang et al. 2004; Li et al. 2011b; Yu et al. 2017b). These findings based on molecular evidence are also supported by recent morphological and cytogenetics studies (Gunathilake et al. 2015; Hembree et al. 2019).

519 The relationships among remaining genera within Theaceae were generally consistent with a
520 recent study using the same molecular markers (Yu et al. 2017b). Within *Camellia*, the backbone
521 remained poorly resolved, most likely reflecting numerous hybridizations and polyploidization
522 events during the rapid radiation of this clade (Yang et al. 2013). *Camellia* species only occur within
523 the Sino-Japanese region, so the uncertainty in the phylogenetic placement of different sections does
524 not affect our biogeographical results.

525 *Northern Hemisphere Origin, Dispersal, and Extinction*

526 Including fossils in the biogeographical analysis revealed a broad mid- to high-latitude
527 Northern Hemispheric origin of Theaceae, strongly supporting the boreotropical forest hypothesis.
528 The different ancestral area reconstructions all inferred Eurasia to be part of the ancestral distribution
529 for the basal node, in most model outcomes also including the Nearctic. Furthermore, the crown age
530 was estimated to be 66.4 Ma (60.0-73.4 Ma), well within the time frame where boreotropical forest
531 is hypothesized to have been widespread and continuous (Tiffney 1985a, 1985b; Lavin and Luckow
532 1993). The reconstruction revealed more range expansion and extinction events compared to analysis
533 based only on extant taxa, and especially captured the expansion to Sino-Japanese from Eurasian
534 after the late Eocene and the extinction in Eurasian around E-O boundary as well as mid to late
535 Miocene.

536 In further support of the boreotropical hypothesis, most reconstructed dispersal events were
537 from the Old to the New World and from north to south (i.e., Eurasian to Sino-Japanese, Sino-
538 Japanese to Papua-Melanesian in the Old World; Nearctic to Panamanian in the New World) (Fig.
539 3). This supports the intriguing notion that there may be a single historical cause of this distribution
540 pattern across many taxa. Specifically, the two tribes that included disjunctive NS distribution (i.e.,
541 Gordonieae and Stewartieae) were both estimated to have undergone vicariance between Eurasia and
542 Nearctic, followed by dispersal to Sino-Japanese from Eurasian and subsequently the extinction in
543 Eurasian within the late Oligocene to mid (late)-Miocene (Fig. 3).

The BioGeoBEARS model allowed species to disperse between the Old and the New World along two different pathways, through the North Atlantic Bridge and the Bering Land Bridge. In the Paleocene and Eocene, species may have taken advantage of both routes during warm intervals (Brikiatis 2014; Wen et al. 2016) though the North Atlantic Land Bridge is considered most likely in the Eocene (Tiffney and Manchester 2002; Brikiatis 2014). Although it was hypothesised that climatic cooling after the late Eocene (approximately 35Ma) made intercontinental dispersal for thermophilic lineages less unlikely (Tiffney and Manchester 2002), this predates the divergence time between extant Old and New World lineages for all clades in the tea family. Gene flow between the continents must have persisted during the Oligocene and possibly until the mid-Miocene, which was observed in several other Northern Hemisphere lineages and attributed to the Bering Land Bridge (Donoghue and Smith 2004; Manos and Meireles 2015).

The reasons for a sustained gene flow between the continents until the mid-Miocene could either be that the changed environment was still within the tolerance limits of the contemporary species, or, the ancestral group managed to develop traits that are better adapted to cooler environments and were able to disperse over the sea in the late Cenozoic. In *Gordonia*, the gene flow around the Oligocene-Miocene boundary supports that it once inhabited mixed mesophytic forest in the western North America in the Miocene which also agrees with paleobotanical evidence (Fig. 3, Baskin and Baskin 2016). Notably, two deciduous species (i.e, *Franklinia alatamaha* and *Stewartia malacodendron*) were involved in the disjunctive patterns, suggesting that some species may have evolved cold tolerance before the mid-Oligocene (Tiffney 1985b), as the oldest deciduous species *Stewartia malacodendron* was dated at 15.1-28.7Ma. Thus, it is possible that some temperate disjunctions might have a tropical origin. This view is also supported by the ancestral state reconstruction analysis of Yu et al. (2017b), in which they recovered the crown Theaceae as an evergreen species. A recent study of clusioid Malpighiales using direct paleoclimate simulation methods shows similar results and indicates that some boreotropical descendants might persist

through niche evolution toward temperate climate (Meseguer et al. 2018). It may have been still possible for temperate plant taxa to cross the Atlantic via the North Atlantic Land Bridge in late Miocene (Denk et al. 2010; Brikiatis 2014).

Under the current phylogenetic hypothesis, dating scheme and fossil evidence, the occurrence of *Laplacea* in Central America can only be explained by a long-distance dispersal event across the Pacific Ocean from the Sino-Japanese region around the mid-Oligocene. *Laplacea*'s sister clade is *Apterosperma* with a single species in eastern China. No fossils relate to these two groups are known. Nevertheless, we cannot rule out the boreotropical hypothesis for this disjunction, given their absence on intervening Pacific islands. Moreover, we infer several dispersals from north to south almost in the same time period of high latitude extinction (Fig. 4b), in line with observations in other clades with amphi-Pacific distribution (Thomas et al. 2017; Yang et al. 2017). One dispersal from Nearctic to Panamanian, possibly occurred along the branch of *Gordonia brangeei*. Another one is from Sino-Japanese to as far as Papua-Melanesian region, possibly within Theeae. The group crossed the Wallace line and supports the view that Southeast Asian biodiversity includes immigrants from northwestern regions such as Indochina (De Bruyn et al. 2014). Increasing range occupancy setting from two to three found similar results (Table S11).

The Importance and Uncertainties of Using Fossils in Biogeographical Analysis

One of the most striking results of this analysis is the failure of the biogeographic reconstruction to infer a realistic clade history in the absence of incorporating fossil evidence. The importance of fossils for estimating ancestral range dynamics is increasingly becoming clear, and a growing number of studies have taken a variety of approaches to incorporate them (Crisp et al. 2011; Meseguer et al. 2015; Sanmartín and Meseguer 2016a). Yet, the great majority of currently published studies lack fossil information, a potentially concerning situation given our results.

Here we used a sequential inference method that focuses more narrowly on evaluating the impact of adding fossils on the ancestral states of key nodes, and thus allows comparing inferences

derived from different sources of data. We made the most conservative assumptions for the fossils by attaching them randomly to the phylogeny within the limits set by well-established taxonomy assignment and age constraints. We then took an empirical-Bayesian method to estimate ancestral states across the distribution of the resulting trees. The method is not computationally intensive, which allows its use on large phylogenies like that of the tea family, or even larger. We show that including even minimal fossil distribution information in biogeographical analysis may lead to substantially different results. The method uncovered a complex historical range dynamic of the Theaceae governed by past environmental change and niche evolution.

To further explore the impact of uncertain phylogenetic placement of fossils and the performance of our method, we compared our approach with an approach recently developed by Landis et al. (2020). The Landis et al. (2020) method parametrically integrates different sources of uncertainty and accounts for their effects on biogeographic reconstructions and temporal estimates using a hierarchical Bayesian framework implemented in RevBayes (Höhna et al., 2016). This method aims to find evolutionary histories that are in harmony across all available lines of evidence, effectively fitting the inferred history as closely to the data as possible. Though potentially very powerful, the computational complexity of this approach makes it too computationally demanding to be applied to datasets as large as our full dataset, even on modern centralized computing clusters. Therefore, we applied the method of Landis et al. (2020) and our method to a subset of taxa from our dataset that included the members of Stewartieae (with generous assistance of Michael Landis, pers. comm, Supplementary Note S1). Both methods produced congruent results under the M1 time-stratified dispersal scenario, but differed slightly under the M0 dispersal scenario (Supplementary Note S1). The comparison shows that the greatest advantage of including fossil data in biogeographical analysis is that they constrain ancestral states of certain clades (Landis et al. 2020), with clear effects even though only a small portion of the recorded fossils have published justifications for relationships with extant species and thus may be used in analyses. Unfortunately,

we cannot confidently infer the southern boundary of the family distribution during the Paleocene and Eocene, as a large portion of the sampled trees inferred the crown group occurrences to include the Sino-Japanese region when using the dataset with 22 fossil taxa (Table S10-S11).

In addition, we tested manually minimizing state space in the biogeographic reconstructions on the extant species phylogeny allowing only single or adjacent regions. We recovered similar marginal probability for different states of key nodes as that of analysis including 10 fossils (Supplementary Fig. S10), supporting the boreotropical hypothesis. However, when we conducted BSM to simulate the range evolution process, the realizations failed with a warning of complex history on a branch and disallowed necessary intermediate states. This test again stresses the importance of increasing sampling over setting stricter priors in modeling biogeographic processes.

Ideally, fossils provide information on traits, time, and distribution. Apart from directly incorporating fossil distributions into parametric biogeographic models, other methods could be explored to improve the understanding of movements between regions. For example, using fossil occurrences to model the historical niche of clades (e.g., Meseguer et al. 2015).

Conclusions

Our study provides insights into the origin of the amphi-Pacific disjunctive distribution of Theaceae and its historical latitudinal range dynamics. We argue for the importance of incorporating fossil information in phylogeny-based biogeographical analysis and we provide a novel method to incorporate such information in biogeographic analyses. We show that even randomly associating fossils to extant phylogenies using limited constraints such as age and genus-level taxonomic information appears to lead to more accurate results for clades where extinction rate was high or spatially biased. Although inference of extinction and dispersal rates was not strongly influenced by including fossils in the biogeographic parametric models, this allowed us to demonstrate repetitive expansion and extraction of the putative distribution of the tea family from mid to high latitude regions in the Northern Hemisphere.

SUPPLEMENTARY MATERIAL

Supplementary material can be found in the Dryad data repository: https://doi.org/10.5061/dryad.x0k6djhh0.

FUNDING

This work was supported by the Danish National Research Foundation (DNRF96 to Y.Y., C. R., and M. K. B.); the Chinese Scholarship Council (No. 201606010394 to Y.Y.); the Norwegian Metacenter for Computational Science (NOTUR; project NN9601K to D.D); Harvard University (the setup funding to C.C.D), and a Carlsberg Young Researcher Award (CF19-0695 to MKB).

ACKNOWLEDGEMENTS

We thank the Harvard University Herbaria and the New York Botanical Garden who generously provided the material for the study. We thank the Bauer Core Facility of the Harvard University for providing technical support during the laboratory process. We thank Sen Li, Petter Marki, Liming Cai, Elizabeth Spriggs, Xiaoshan Duan and Camille Desisto for helping with bioinformatics and wet lab work. We thank three anonymous reviewers for extremely insightful comments on the manuscript. The computations in this paper were partly run on the FASRC Odyssey cluster supported by the FAS Division of Science Research Computing Group at Harvard University and partly run on the CIPRES Science Gateway.

REFERENCES

Aberer A.J., Krompass D., Stamatakis A. 2013. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. Syst. Biol. 62:162–6.

Alamoudi E.F., Khalil W.K.B., Ghaly I.S., Hassan N.H.A., Ahmed E.S. 2014. Nanoparticles from of costus speciosus extract improves the antidiabetic and antilipidemic effects against STZ-induced diabetes mellitus in albino rats. Int. J. Pharm. Sci. Rev. Res. 29:279–288.

Antonelli A., Nylander J.A.A., Persson C., Sanmartin I. 2009. Tracing the impact of the Andean

668 uplift on Neotropical plant evolution. Proc. Natl. Acad. Sci. 106:9749–9754.

669 Baskin J.M., Baskin C.C. 2016. Origins and Relationships of the Mixed Mesophytic Forest of

670 Oregon–Idaho, China, and Kentucky: Review and Synthesis [1]. Ann. Missouri Bot. Gard.

671 101:525–552.

672 Beaulieu J.M., Tank D.C., Donoghue M.J. 2013. A Southern Hemisphere origin for campanulid

673 angiosperms, with traces of the break-up of Gondwana. BMC Evol. Biol. 13:80.

674 Bouckaert R., Heled J., Kühnert D., Vaughan T., Wu C.-H., Xie D., Suchard M.A., Rambaut A.,

675 Drummond A.J. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis.

676 PLoS Comput. Biol. 10:e1003537.

677 Bozukov V., Palamarev E. 1995. On the Tertiary history of the Theaceae in Bulgaria. Flora Mediterr.

678 5:177–190.

679 Brikiatis L. 2014. The de geer, thulean and beringia routes: Key concepts for understanding early

680 cenozoic biogeography. J. Biogeogr. 41:1036–1054.

681 De Bruyn M., Stelbrink B., Morley R.J., Hall R., Carvalho G.R., Cannon C.H., Van Den Bergh G.,

682 Meijaard E., Metcalfe I., Boitani L., Maiorano L., Shoup R., Von Rintelen T. 2014. Borneo and

683 Indochina are major evolutionary hotspots for Southeast Asian biodiversity. Syst. Biol. 63:879–

684 901.

685 Cai L., Xi Z., Peterson K., Rushworth C., Beaulieu J., Davis C.C. 2016. Phylogeny of Elatinaceae

686 and the tropical Gondwanan origin of the Centroplacaceae (Malpighiaceae, Elatinaceae) clade.

687 PLoS One. 11:1–21.

688 Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden T.L. 2009.

689 BLAST+: architecture and applications. BMC Bioinformatics. 10:421.

690 Capella-Gutiérrez S., Silla-Martínez J.M., Gabaldón T. 2009. trimAl: A tool for automated alignment

691 trimming in large-scale phylogenetic analyses. Bioinformatics. 25:1972–1973.

692 Chacón J., Renner S.S. 2014. Assessing model sensitivity in ancestral area reconstruction using

693 Lagrange: A case study using the Colchicaceae family. J. Biogeogr. 41:1414–1427.
694 Chen S., Zhou Y., Chen Y., Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor.
695 Bioinformatics. 34:i884–i890.
696 Christenhusz M.J.M., Chase M.W. 2013. Biogeographical patterns of plants in the Neotropics -
697 dispersal rather than plate tectonics is most explanatory. Bot. J. Linn. Soc. 171:277–286.
698 Condamine F.L., Sperling F.A.H., Kergoat G.J. 2013. Global biogeographical pattern of swallowtail
699 diversification demonstrates alternative colonization routes in the Northern and Southern
700 hemispheres. J. Biogeogr. 40:9–23.
701 Crisp M.D., Trewick S.A., Cook L.G. 2011. Hypothesis testing in biogeography. Trends Ecol. Evol.
702 26:66–72.
703 Davis C.C., Bell C.D., Mathews S., Donoghue M.J. 2002. Laurasian migration explains Gondwanan
704 disjunctions: evidence from Malpighiaceae. Proc Natl Acad Sci U S A. 99.
705 Denk T., Grímsson F., Zetter R. 2010. Episodic migration of oaks to Iceland: Evidence for a north
706 atlantic “land bridge” in the latest miocene. Am. J. Bot. 97:276–287.
707 Dodsworth S. 2015. Genome skimming for next-generation biodiversity analysis. Trends Plant Sci.
708 20:525–527.
709 Donoghue M.J., Smith S.A. 2004. Patterns in the assembly of temperate forests around the Northern
710 Hemisphere. Philos. Trans. R. Soc. B Biol. Sci. 359:1633–1644.
711 Dupin J., Matzke N.J., Särkinen T., Knapp S., Olmstead R.G., Bohs L., Smith S.D. 2017. Bayesian
712 estimation of the global biogeographical history of the Solanaceae. J. Biogeogr. 44:887–899.
713 Fritsch P.W., Manchester S.R., Stone R.D., Cruz B.C., Almeda F. 2015. Northern Hemisphere
714 origins of the amphi-Pacific tropical plant family Symplocaceae. J. Biogeogr. 42:891–901.
715 Grote P., Dilcher D. 1989. Investigations of angiosperms from the Eocene of North America: A new
716 genus of Theaceae based on fruit and seed remains. Bot. Gaz. 150:190–206.
717 Grote P.J.., Dilcher D.L.. 1992. Fruits and Seeds of Tribe Gordonieae (Theaceae) from the Eocene of

718 North America. Am. J. Bot. 79:744–753.
719 Gunathilake L.A.A.H., Prince J.S., Whitlock B.A. 2015. Seed coat micromorphology of Gordonia
720 sensu lato (including Polyspora and Laplacea; Theaceae). Brittonia. 67:68–78.
721 Hembree W.G., Ranney T.G., Jackson B.E., Weathington M. 2019. Cytogenetics, ploidy, and
722 genome sizes of Camellia and related genera. HortScience. 54:1124–1142.
723 Holt B.G., Lessard J.-P., Borregaard M.K., Fritz S.A., Araujo M.B., Dimitrov D., Fabre P.-H.,
724 Graham C.H., Graves G.R., Jonsson K.A., Nogues-Bravo D., Wang Z., Whittaker R.J., Fjeldsa
725 J., Rahbek C. 2013. An Update of Wallace's Zoogeographic Regions of the World. Science
726 (80-.). 339:74–78.
727 Höhna, Landis, Heath, Boussau, Lartillot, Moore, Huelsenbeck, Ronquist. 2016. RevBayes:
728 Bayesian phylogenetic inference using graphical models and an interactive model-specification
729 language. Systematic Biology. 65:726-736.
730 Huelsenbeck J.P., Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees.
731 Bioinformatics. 17:754–755.
732 Katoh K., Misawa K., Kuma K., Miyata T. 2002. MAFFT: a novel method for rapid multiple
733 sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–66.
734 Kozlov A.M., Darriba D., Flouri T., Morel B., Stamatakis A. 2019. RAxML-NG: a fast, scalable and
735 user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics.
736 Landis M.J., Eaton D.A.R., Clement W.L., Park B., Spriggs E.L., Sweeney P.W., Edwards E.J.,
737 Donoghue M.J. 2020. Joint Phylogenetic Estimation of Geographic Movements and Biome
738 Shifts during the Global Diversification of Viburnum. Syst. Biol. 70:67–85.
739 Lanfear R., Frandsen P.B., Wright A.M., Senfeld T., Calcott B. 2016. PartitionFinder 2: New
740 Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological
741 Phylogenetic Analyses. Mol. Biol. Evol.:msw260.
742 Lavin M., Luckow M. 1993. Origins and relationships of tropical North America in the context of

743 the boreotropics hypothesis. Am. J. Bot. 80:1–14.

744 Li J., Del Tredici P., Yang S., Donoghue M.J. 2002. Phylogenetic relationships and biogeography of

745 *Stewartia* (Camellioideae, Theaceae) inferred from nuclear ribosomal DNA ITS sequences.

746 Rhodora. 104:117–033.

747 Li L., Li J., Rohwer J.G., van der Werff H., Wang Z.H., Li H.W. 2011a. Molecular phylogenetic

748 analysis of the Persea group (Lauraceae) and its biogeographic implications on the evolution of

749 tropical and subtropical Amphi-Pacific disjunctions. Am. J. Bot. 98:1520–1536.

750 Li R., Wen J. 2013. Phylogeny and Biogeography of Dendropanax (Araliaceae), an Amphi-Pacific

751 Disjunct Genus Between Tropical/Subtropical Asia and the Neotropics. Syst. Bot. 38:536–551.

752 Li R., Yang J.B., Yang S.X., Li D.Z. 2011b. Phylogeny and taxonomy of the Pyrenaria complex

753 (Theaceae) based on nuclear ribosomal ITS sequences. Nord. J. Bot. 29:780–787.

754 Li Y., Awasthi N., Yang J., Li C. Sen. 2013. Fruits of Schima (Theaceae) and seeds of Toddalia

755 (Rutaceae) from the Miocene of Yunnan Province, China. Rev. Palaeobot. Palynol. 193:119–

756 127.

757 Lin H.-Y., Hao Y.-J., Li J.-H., Fu C.-X., Soltis P.S., Soltis D.E., Zhao Y.-P. 2019. Phylogenomic

758 conflict resulting from ancient introgression following species diversification in Stewartia s.l.

759 (Theaceae). Mol. Phylogenet. Evol. 135:1–11.

760 Mai U., Mirarab S. 2018. TreeShrink: Fast and accurate detection of outlier long branches in

761 collections of phylogenetic trees. BMC Genomics. 19.

762 Manos P.S., Meireles J.E. 2015. Biogeographic analysis of the woody plants of the Southern

763 Appalachians: Implications for the origins of a regional flora. Am. J. Bot. 102:780–804.

764 Mao K., Milne R.I., Zhang L., Peng Y., Liu J., Thomas P., Mill R.R., S. Renner S. 2012. Distribution

765 of living Cupressaceae reflects the breakup of Pangea. Proc. Natl. Acad. Sci. 109:7793–7798.

766 Marinho L.C., Cai L., Duan X., Ruhfel B.R., Fiaschi P., Amorim A.M., van den Berg C., Davis C.C.

767 2019. Plastomes resolve generic limits within tribe Clusieae (Clusiaceae) and reveal the new

768 genus Arawakia. Mol. Phylogenet. Evol. 134:142–151.

769 Matzke N.J. 2013. BioGeoBEARS: BioGeography with Bayesian (and Likelihood) Evolutionary

770 Analysis in R Scripts. .

771 Meseguer A.S., Condamine F.L. 2017a. Ancient tropical extinctions contributed to the latitudinal

772 diversity gradient. bioRxiv. 2:236646.

773 Meseguer A.S., Condamine F.L. 2017b. Ancient tropical extinctions contributed to the latitudinal

774 diversity gradient. bioRXiv. 3.

775 Meseguer A.S., Lobo J.M., Cornuault J., Beerling D., Ruhfel B.R., Davis C.C., Jousselin E.,

776 Sanmartín I. 2018. Reconstructing deep-time palaeoclimate legacies in the clusioid

777 Malpighiales unveils their role in the evolution and extinction of the boreotropical flora. Glob.

778 Ecol. Biogeogr. 27:616–628.

779 Meseguer A.S., Lobo J.M., Ree R., Beerling D.J., Sanmartín I. 2015. Integrating fossils,

780 phylogenies, and niche models into biogeography to reveal ancient evolutionary history: The

781 case of Hypericum (Hypericaceae). Syst. Biol. 64:215–232.

782 Miller M.A., Pfeiffer W., Schwartz T. 2010. Creating the CIPRES Science Gateway for inference of

783 large phylogenetic trees. 2010 Gatew. Comput. Environ. Work. GCE 2010.

784 Min T., Bartholomew B. 2007. Theaceae. Flora of China. p. 366–478.

785 Nauheimer L., Metzler D., Renner S.S. 2012. Global history of the ancient monocot family Araceae

786 inferred with models. New Phytol. 195:938–950.

787 Nylander J.A.A., Olsson U., ALSTRöM P., Sanmartín I. 2008. Accounting for phylogenetic

788 uncertainty in biogeography: A bayesian approach to dispersal-vicariance analysis of the

789 thrushes (Aves: Turdus). Syst. Biol. 57:257–268.

790 Price M.N., Dehal P.S., Arkin A.P. 2010. FastTree 2 – Approximately Maximum-Likelihood Trees

791 for Large Alignments. PLoS One. 5:e9490.

792 Prince L. 2007. A Brief Nomenclatural Review of Genera and Tribes in Theaceae. Aliso. 24:105–

121.

Prince L.M. 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 L.M., Parks C.R. 2001. Phylogenetic relationships of Theaceae inferred from chloroplast DNA sequence data. Am. J. Bot. 88:2309–2320.

Rao M., Steinbauer M.J., Xiang X., Zhang M., Mi X., Zhang J., Ma K., Svenning J.C. 2018. Environmental and evolutionary drivers of diversity patterns in the tea family (Theaceae s.s.) across China. Ecol. Evol.:11663–11676.

Ree R.H., Sanmartín I. 2018. Conceptual and statistical problems with the DEC+J model of founder-event speciation and its comparison with DEC via model selection. J. Biogeogr. 45:741–749.

Ree R.H., Smith S.A. 2008. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. Syst. Biol. 57:4–14.

Ripma L.A., Simpson M.G., Hasenstab-Lehman K. 2014. Geneious! Simplified Genome Skimming Methods for Phylogenetic Systematic Studies: A Case Study in *Oreocarya* (Boraginaceae). Appl. Plant Sci. 2:1400062.

Rose J.P., Kleist T.J., Löfstrand S.D., Drew B.T., Schönenberger J., Sytsma K.J. 2018. Phylogeny, historical biogeography, and diversification of angiosperm order Ericales suggest ancient Neotropical and East Asian connections. Mol. Phylogenet. Evol. 122:59–79.

Sanmartín I., Enghoff H., Ronquist F. 2001. Patterns of animal dispersal, vicariance and diversification in the Holarctic. Biol. J. Linn. Soc. 73:345–390.

Sanmartín I., Meseguer A.S. 2016a. Extinction in phylogenetics and biogeography: From timetrees to patterns of biotic assemblage. Front. Genet. 7:1–17.

Sanmartín I., Meseguer A.S. 2016b. Extinction in phylogenetics and biogeography: From timetrees to patterns of biotic assemblage. Front. Genet. 7:1–17.

818 Smith S.A. 2009. Taking into account phylogenetic and divergence-time uncertainty in a parametric
819 biogeographical analysis of the Northern Hemisphere plant clade Caprifolieae. J. Biogeogr.
820 36:2324–2337.
821 Smith S.A., O'Meara B.C. 2012. treePL: divergence time estimation using penalized likelihood for
822 large phylogenies. Bioinformatics. 28:2689–2690.
823 Steenis C.G.G.J. van. 1962. The land-bridge theory in botany with particular reference to tropical
824 plants. Blumea - Tijdschr. voor Syst. en Geogr. der Planten. 11:235–372.
825 Thomas D.C., Tang C.C., Saunders R.M.K. 2017. Historical biogeography of Goniothalamus and
826 Annonaceae tribe Annoneae: dispersal–vicariance patterns in tropical Asia and intercontinental
827 tropical disjunctions revisited. J. Biogeogr. 44:2862–2876.
828 Thorne R. 1972. Major Disjunctions in the Geographic Ranges of Seed Plants. Q. Rev. Biol. 90:365–
829 411.
830 Tiffney B. 1985a. The Eocene North Atlantic Land Bridge: its importance in Teritiary and modern
831 phytogeography of the Northern Hemisphere. J. Arnold Arbor. 66:243–273.
832 Tiffney B.H. 1985b. Perspectives on the origin of the floristic similarity between Eastern Asia and
833 Eastern North America. J. Arnold Arboretum. 66:73–94.
834 Tiffney B.H., Manchester S.R. 2002. The Use of Geological and Paleontological Evidence in
835 Evaluating Plant Phylogeographic Hypotheses in the Northern Hemisphere Tertiary. Int. J. Plant
836 Sci. 162:S3–S17.
837 Walker D., Geissman J.., Compilers. 2018. GSA Geologic time scale v. 5.0. Geol. Soc. Am.
838 204:59425.
839 Wen J., Ickert-Bond S., Nie Z.-L., Li R. 2010. Timing and Modes of Evolution of Eastern Asian -
840 North American Biogeographic Disjunctions in Seed Plants. Darwin's Herit. Today Proc.
841 Darwin 2010 Beijing Int. Conf.:252–269.
842 Wen J., Nie Z.-L., Ickert-Bond S.M. 2016. Intercontinental disjunctions between eastern Asia and

western North America in vascular plants highlight the biogeographic importance of the Bering land bridge from late Cretaceous to Neogene. J. Syst. Evol. 54:469–490.

Wood H.M., Matzke N.J., Gillespie R.G., Griswold C.E. 2013. Treating fossils as terminal taxa in divergence time estimation reveals ancient vicariance patterns in the palpimanoid spiders. Syst. Biol. 62:264–284.

Wu Z.Y., Liu J., Provan J., Wang H., Chen C.J., Cadotte M.W., Luo Y.H., Amorim B.S., Li D.Z., Milne R.I. 2018. Testing Darwin's transoceanic dispersal hypothesis for the inland nettle family (Urticaceae). Ecol. Lett. 21:1515–1529.

Xiang X.G., Mi X.C., Zhou H.L., Li J.W., Chung S.W., Li D.Z., Huang W.C., Jin W.T., Li Z.Y., Huang L.Q., Jin X.H. 2016. Biogeographical diversification of mainland Asian Dendrobium (Orchidaceae) and its implications for the historical dynamics of evergreen broad-leaved forests. J. Biogeogr. 43:1310–1323.

Yang J.B., Yang S.X., Li H.T., Yang J., Li D.Z. 2013. Comparative Chloroplast Genomes of Camellia Species. PLoS One. 8.

Yang M.Q., Li D.Z., Wen J., Yi T.S. 2017. Phylogeny and biogeography of the amphi-Pacific genus Aphananthe. PLoS One. 12:1–18.

Yang S.X., Yang J.B., Lei L.G., Li D.Z., 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 Syst. Evol. 248:45–55.

Yu G., Smith D.K., Zhu H., Guan Y., Lam T.T.Y. 2017a. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. 8:28–36.

Yu X.Q., Gao L.M., Soltis D.E., Soltis P.S., Yang J.B., Fang L., Yang S.X., Li D.Z. 2017b. Insights into the historical assembly of East Asian subtropical evergreen broadleaved forests revealed by the temporal history of the tea family. New Phytol. 215:1235–1248.

Zeng C.X., Hollingsworth P.M., Yang J., He Z.S., Zhang Z.R., Li D.Z. 2018. Genome skimming herbarium specimens for DNA barcoding and phylogenomics. Plant Methods.:1–14.

Zizka A., Silvestro D., Andermann T., Azevedo J., Duarte Ritter C., Edler D., Farooq H., Herdean A., Ariza M., Scharn R., Svantesson S., Wengström N., Zizka V., Antonelli A. 2019. CoordinateCleaner : Standardized cleaning of occurrence records from biological collection databases. Methods Ecol. Evol. 10:744–751.

900

901

902 FIGURE CAPTIONS

903 **Figure 1**. Current richness distribution pattern of Theaceae and occurrences of fossils through time
904 and across biogeographic regions. The stack plot shows the number of reported fossils in each
905 biogeographic region during different epoch. The capital letters on the map represent different
906 biogeographic regions defined in the study. N: Nearctic, P: Panamanian, E: Eurasian, S: Sino-
907 Japanese, M: Papua-Melanesia.

908 **Figure 2**. Left: Phylogeny of Theaceae inferred using the combined pt-nrDNA dataset in RAxML-
909 ng. Numbers associated with nodes indicate ML bootstrap support (BS) values. Asterisks represent
910 nodes with BS values >90%. Right: The time-calibrated phylogeny of Theaceae. Topology is derived
911 from the maximum likelihood tree on the left. Two species of Lecythidaceae at the root of the
912 phylogeny are pruned as the root height is set manually in TreePL and does not influence the height
913 of inner nodes in our study. 95% confidence intervals of the divergence time estimated in treePL are
914 shown as blue bars at each node. Fossil calibrations are marked by orange circles with numbers
915 corresponding to Table S5. The time intervals annotated below the phylogeny corresponds to the six
916 geological epochs in Figure 1. Photos above the branches of the three tribes show the flower of
917 *Camellia oleifera*, *Franklinia alatamaha*, *Stewartia pseudocamellia* from top to bottom.

918 **Figure 3**. Biogeographic reconstruction of Theaceae showing the effect of incorporating fossil
919 information into parametric biogeographic models. The reconstruction used DEC model under M0
920 dispersal scenario with maximum number of occupied ranges set to two over 300 phylogenies. a-b.
921 Average marginal probability of different areas and the biogeographic stochastic mapping (BSM)
922 result for majority of trees mapped on the time-calibrated best RAxML-ng tree based on only extant
923 taxa (Fig.3a) or based on both extant taxa and 10 fossil taxa (Fig.3b). Colored circles at tips represent

current ranges. Pie charts at inner nodes represent the average marginal probability of ranges across the 300 phylogenies and probabilities lower than 0.1 were combined and shown in white. The bar plots show the distribution of BSM result of the crown Theaceae across 300 phylogenies. Colored squares represent the recovered ranges of key nodes that have the highest frequencies across 300 phylogenies. Arrows and red crosses annotate the dispersal and extinction events that generated the distribution patterns. Periods of forest types are divided following Meseguer et al. (2015). c-d. Average number of reconstructed dispersals between different regions using only extant taxa (Fig. 3c) or using both extant and 10 extinct taxa (Fig. 3d). One tick mark represents one dispersal event. The lower panel show the source ranges and the upper panel show the sink ranges. The colored band within the lower panel show the sink ranges for each source range. The reconstructions using different model settings are presented in Supplementary Fig. S5-S9.

Figure 4. Estimated number of events through time based on biogeographic models using only extant taxa (a) and models incorporating fossil information (b). The results are the average numbers of 300 phylogenies with error bars indicate standard deviations. Shading shows the hypothesized three die out stages of the mid- to high- latitude boreotropical forest corridor followed Meseguer et al. 2015. “Trans.” represents mixed mesophytic forest and “Cold” represents boreal and temperate forest.

941

TABLE 1. Summary of ancestral range estimation using biogeographic models with maximum number of occupied ranges set to two (mean ± standard deviation).

Dataset	Model	LnL	num_params	d	e	j	AICc	AICc_wt
nfossil	M0-DEC	-64.012±2.741	2	0.004±0.001	0±0	0±0	132.121±5.482	0.185
	M0-DEC+j	-61.199±1.884	3	0.002±0.001	0±0	0.01±0.003	128.591±3.769	0.815
	M1-DEC	-63.773±2.73	2	0.004±0.001	0±0	0±0	131.641±5.459	0.195
	M1-DEC+j	-61.046±1.887	3	0.002±0.001	0±0	0.01±0.003	128.285±3.774	0.805
fossil_10	M0-DEC	-103.367±4.418	2	0.008±0.001	0.002±0.001	0±0	210.823±8.837	0.010
	M0-DEC+j	-86.777±4.288	3	0.002±0.001	0±0	0.045±0.006	179.733±8.576	0.990
	M1-DEC	-104.312±4.406	2	0.008±0.001	0.001±0.001	0±0	212.713±8.812	<0.001
	M1-DEC+j	-88.358±3.453	3	0.002±0	0±0	0.045±0.004	182.895±6.906	>0.999
fossil_22	M0-DEC	-129.707±5.347	2	0.01±0.002	0.003±0.001	0±0	263.495±10.695	<0.001

M0-DEC+j	-109.218±6.066	3	0.002±0.001	0±0	0.055±0.004	224.601±12.132	>0.999
M1-DEC	-131.21±5.422	2	0.01±0.002	0.003±0.001	0±0	266.501±10.844	<0.001
M1-DEC+j	-110.292±5.915	3	0.002±0.001	0±0	0.057±0.004	226.749±11.831	>0.999

Notes: LnL = log value of the likelihood; d = rate of dispersal; e = rate of extinction; j = relative per-event weight of jump dispersal

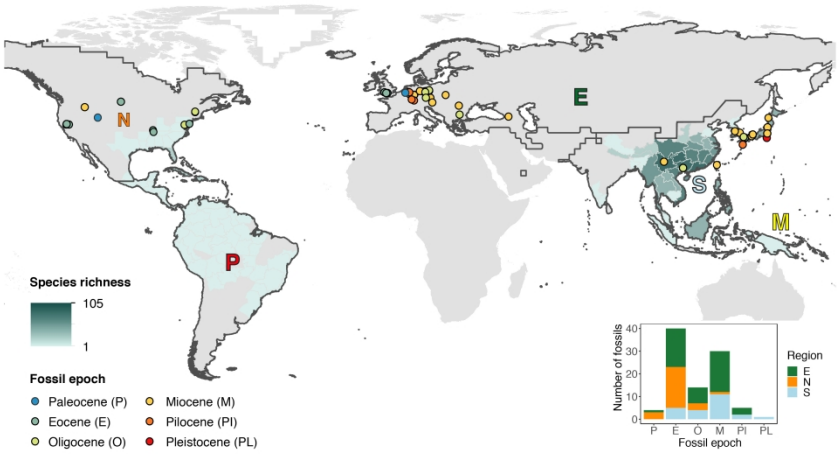

Figure 1. Current richness distribution pattern of Theaceae and occurrences of fossils through time and across biogeographic regions. The stack plot shows the number of reported fossils in each biogeographic region during different epoch. The capital letters on the map represent different biogeographic regions defined in the study. N: Nearctic, P: Panamanian, E: Eurasian, S: Sino-Japanese, M: Papua-Melanesia.

341x209mm (300 x 300 DPI)

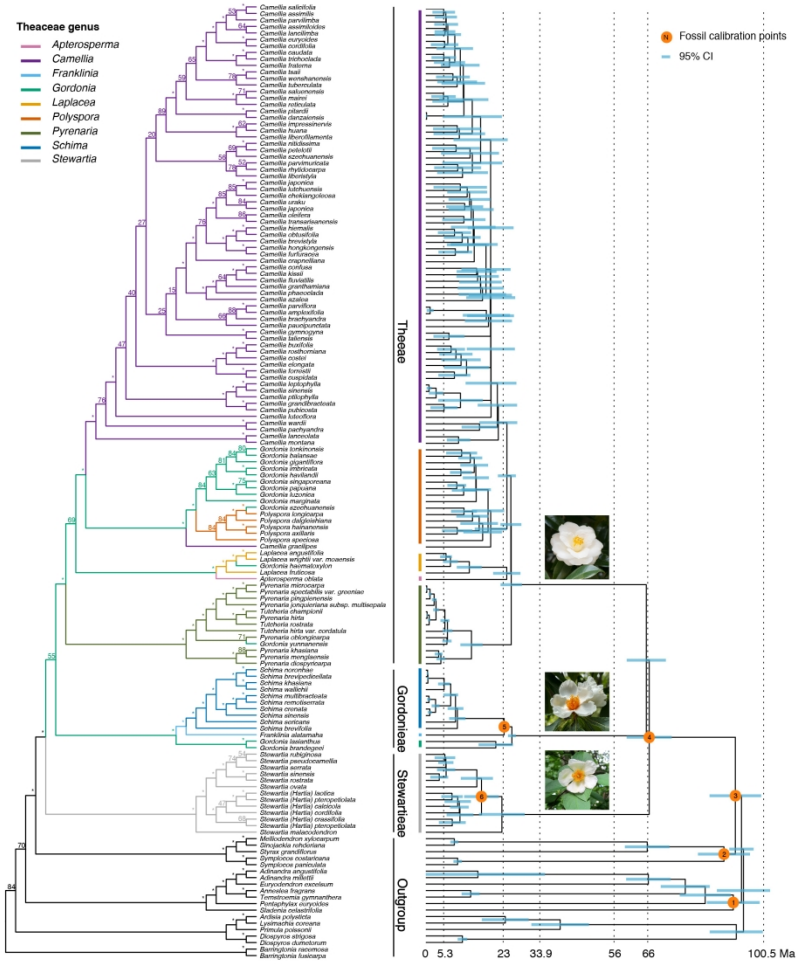

Figure 2. Left: Phylogeny of Theaceae inferred using the combined pt-nrDNA dataset in RAxML-ng. Numbers associated with nodes indicate ML bootstrap support (BS) values. Asterisks represent nodes with BS values >90%. Right: The time-calibrated phylogeny of Theaceae. Topology is derived from the maximum likelihood tree on the left. Two species of Lecythidaceae at the root of the phylogeny are pruned as the root height is set manually in TreePL and does not influence the height of inner nodes in our study. 95% confidence intervals of the divergence time estimated in treePL are shown as blue bars at each node. Fossil calibrations are marked by orange circles with numbers corresponding to Table S5. The time intervals annotated below the phylogeny corresponds to the six geological epochs in Figure 1. Photos above the branches of the three tribes show the flower of Camellia oleifera, Franklinia alatamaha, Stewartia pseudocamellia from top to bottom.

210x296mm (300 x 300 DPI)

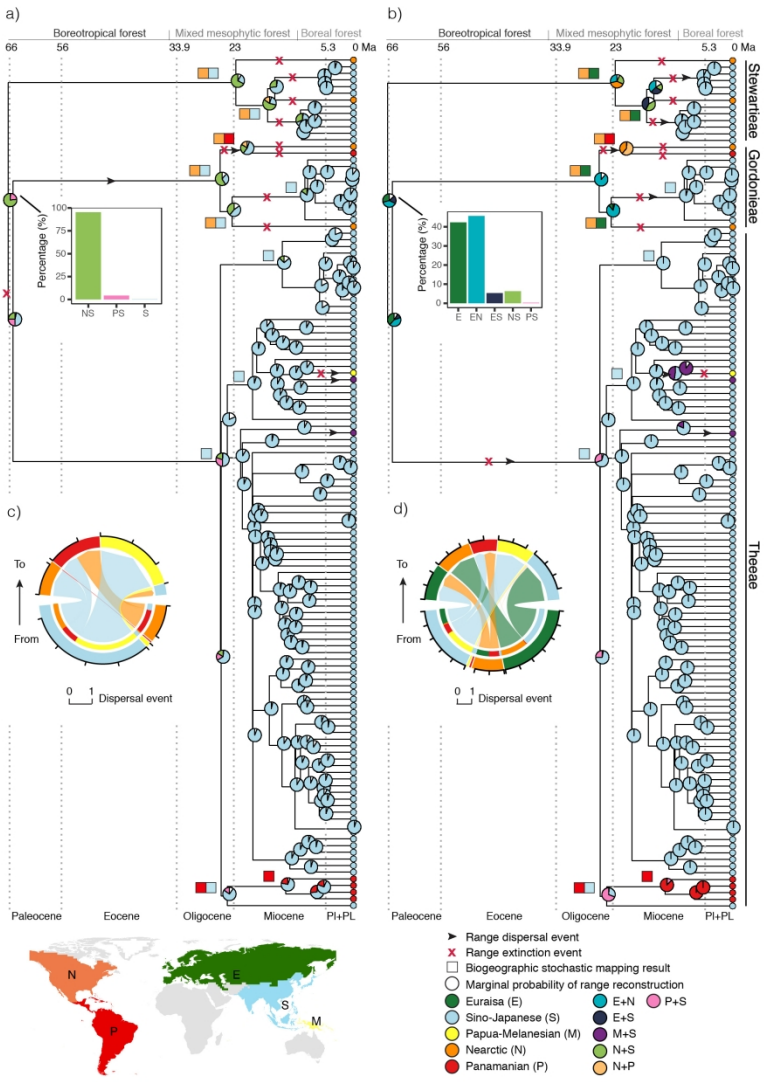

Figure 3. Biogeographic reconstruction of Theaceae showing the effect of incorporating fossil information into parametric biogeographic models. The reconstruction used DEC model under M0 dispersal scenario with maximum number of occupied ranges set to two over 300 phylogenies. a-b. Average marginal probability of different areas and the biogeographic stochastic mapping (BSM) result for majority of trees mapped on the time-calibrated best RAxML-ng tree based on only extant taxa (Fig.3a) or based on both extant taxa and 10 fossil taxa (Fig.3b). Colored circles at tips represent current ranges. Pie charts at inner nodes represent the average marginal probability of ranges across the 300 phylogenies and probabilities lower than 0.1 were combined and shown in white. The bar plots show the distribution of BSM result of the crown Theaceae across 300 phylogenies. Colored squares represent the recovered ranges of key nodes that have the highest frequencies across 300 phylogenies. Arrows and red crosses annotate the dispersal and extinction events that generated the distribution patterns. Periods of forest types are divided following Meseguer et al. (2015). c-d. Average number of reconstructed dispersals between different regions using only extant taxa (Fig. 3c) or using both extant and 10 extinct taxa (Fig. 3d). One tick mark represents one dispersal event. The lower panel show the source ranges and the upper panel show the sink ranges. The colored band within the lower

panel show the sink ranges for each source range. The reconstructions using different model settings are presented in Supplementary Fig. S5-S9.

210x297mm (300 x 300 DPI)

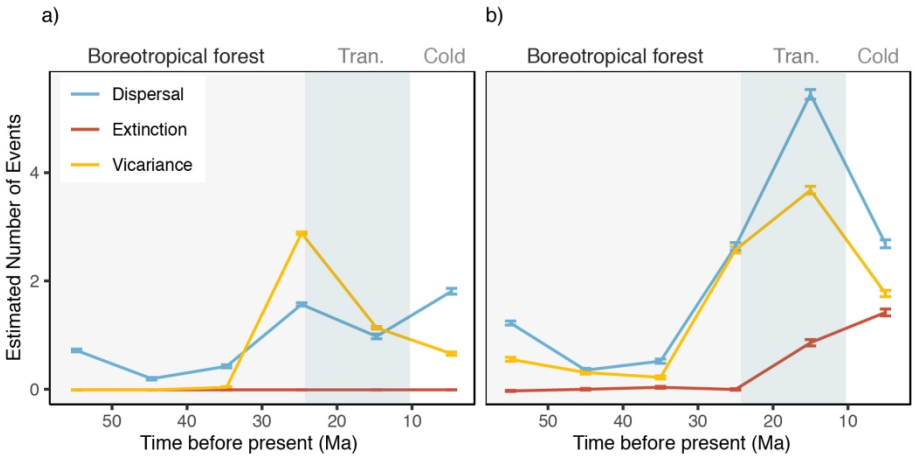

Figure 4. Estimated number of events through time based on biogeographic models using only extant taxa (a) and models incorporating fossil information (b). The results are the average numbers of 300 phylogenies with error bars indicate standard deviations. Shading shows the hypothesized three die out stages of the mid- to high- latitude boreotropical forest corridor followed Meseguer et al. 2015. "Trans." represents mixed mesophytic forest and "Cold" represents boreal and temperate forest.

172x100mm (300 x 300 DPI)