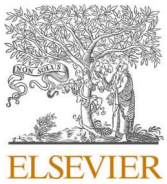
Molecular Phylogenetics and Evolution 196 (2024) 108089



Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

A new nuclear phylogeny of the tea family (Theaceae) unravels rapid radiations in genus *Camellia*

Yujing Yan [a,b,*], Rute R. da Fonseca [a], Carsten Rahbek [a,c,d,e], Michael K. Borregaard [a,#], Charles C. Davis [b,#]

[a] Center for Macroecology, Evolution and Climate, Globe Institute, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen, Denmark
[b] Department of Organismic and Evolutionary Biology, Harvard University Herbaria, 22 Divinity Ave, Cambridge, MA 02138, USA
[c] Center for Global Mountain Biodiversity, Globe Institute, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen, Denmark
[d] Department of Life Sciences, Imperial College London, Silkwood Park campus, Ascot SL5 7PY, UK
[e] Danish Institute for Advanced Study, University of Southern Denmark, 5230 Odense M, Denmark

ARTICLE INFO

Keywords:
Target enrichment
Phylogenetic discordance
Phylogenomics
Herbarium specimen
Incomplete lineage sorting (ILS)
Locus filtering

ABSTRACT

Molecular analyses of rapidly radiating groups often reveal incongruence between gene trees. This mainly results from incomplete lineage sorting, introgression, and gene tree estimation error, which complicate the estimation of phylogenetic relationships. In this study, we reconstruct the phylogeny of Theaceae using 348 nuclear loci from 68 individuals and two outgroup taxa. Sequence data were obtained by target enrichment using the recently released Angiosperm 353 universal probe set applied to herbarium specimens. The robustness of the topologies to variation in data quality was established under a range of different filtering schemes, using both coalescent and concatenation approaches. Our results confirmed most of the previously hypothesized relationships among tribes and genera, while clarifying additional interspecific relationships within the rapidly radiating genus *Camellia*. We recovered a remarkably high degree of gene tree heterogeneity indicative of rapid radiation in the group and observed cytonuclear conflicts, especially within *Camellia*. This was especially pronounced around short branches, which we primarily associate with gene tree estimation error. Our analysis also indicates that incomplete lineage sorting (ILS) contributed to gene-tree conflicts and accounted for approximately 14 % of the explained variation, whereas inferred introgression levels were low. Our study advances the understanding of the evolution of this important plant family and provides guidance on the application of target capture methods and the evaluation of key processes that influence phylogenetic discordances.

1. Introduction

Phylogenomic analyses using hundreds or even thousands of nuclear genes have successfully resolved or provided valuable insights into the phylogenetic relationships of many plants group that were previously difficult to reconstruct using plastid genome or single nuclear genes (e.g., Guo et al., 2020; Leebens-Mack et al., 2019; Zhang et al., 2021). However, analyses of rapidly radiating groups also often reveal discordance between gene trees and species trees (Larson et al., 2020; Léveillé-Bourret et al., 2018; Thomas et al., 2021). Such information is not only useful in clarifying phylogenetic relationships but can also be used to investigate molecular evolutionary processes that contribute to diversification.

Gene tree discordance (variation) can arise from biological factors such as incomplete lineage sorting (ILS) and introgression. ILS allows ancestral genetic polymorphisms to persist during rapid speciation events (Avise and Robinson, 2008) and has complicated phylogenetic inference in many plant lineages (e.g., Cai et al., 2021; Meleshko et al., 2021; Murillo-A et al., 2022). Introgressive hybridization and organelle capture, facilitating gene exchange among taxa, have also been commonly observed at different taxonomic levels during rapid radiations and have contribute to gene tree discordance (e.g., Degnan and Rosenberg, 2009; Suh et al., 2015; Meyer et al., 2017; Muñoz-Rodríguez et al., 2018; Cai et al., 2021). Another source of variation between gene trees is stochastic errors and systematic biases, such as errors in gene tree estimation, which associate with varies factors including the quality

* Corresponding author at: Department of Organismic and Evolutionary Biology, Harvard University Herbaria, 22 Divinity Ave, Cambridge, MA 02138, USA.
E-mail address: yjyan7@gmail.com (Y. Yan).
# These authors have contributed equally to this work.

https://doi.org/10.1016/j.ympev.2024.108089
Received 23 April 2023; Received in revised form 8 March 2024; Accepted 25 April 2024
Available online 27 April 2024

of the data, the analytical methods and model parameters employed in phylogenetic reconstruction. It tends to increase with the decrease in phylogenetic informativeness of the gene (Mirarab et al., 2016; Xi et al., 2015).

Theaceae (the tea family), comprising 372 species classified into nine genera and three tribes (Theeae, Stewartieae and Gordonieae), is mainly distributed in subtropical and tropical Asia. It is notable for its economic and cultural importance, particularly species such as tea (*Camellia sinensis* (L.) Kuntze) and oil plants (e.g., *Camellia oleifera* Abel). Additionally, some lineages of this family are prominent constituents of the subtropical evergreen forests in Asia. Theaeceae is hypothesized to have undergone several rapid diversifications coinciding with the rise of the Eastern Asian Monsoon (Yu et al., 2017; Cheng et al., 2022), contributing to the complexity and richness of the family.

Resolving the phylogenetic relationships within this flowering plant clade has been a long-standing problem, which in recent years has seen a flurry of interest, with multiple phylogenetic revisions employing sequences from nuclear genes (Zhao et al., 2023), chloroplast genomes (Yan et al., 2021; Yu et al., 2017) and even full transcriptomes (Cheng et al., 2022; Zhang et al., 2022). Though these efforts have greatly improved the resolution of intergeneric relationships, studies have inferred inconsistent topologies for the crown of tribe Theeae and within *Camellia,* the central radiations with the most ecologically and culturally important species (over 200 species in *Camellia* alone) (Cheng et al., 2022; Wu et al., 2022; Zan et al., 2023; Zhao et al., 2023). These radiations are speculated to have occurred extremely rapidly around ~ 20 Ma (Cheng et al., 2022; Yan et al., 2021; Zan et al., 2023), coinciding with a climatic shift that triggered the spread of subtropical evergreen forest throughout most of Asia (Yu et al., 2017), holding the key to understand the evolutionary history and ecology of the group.

The difficulty in unravelling the interspecific relationships within these groups date back to the first major taxonomic treatments of the family more than 60 years ago. For *Camellia*, the traditional classification system, based on morphology, was first established by Sealy (1958), who divided the genus into 12 sections, then developed by Chang (1998), who expanded the number of sections to 22. The latest classification by Ming (1999), later adopted in the Flora of China by Min and Bartholomew (2007), simplified the taxonomy to include only 14 sections. Each successive revision saw marked changes in both section boundaries and hypothesized inter-section relationships. More recent large-scale attempts to resolve the genus using molecular markers, including RNA polymerase II (RPB2), the ITS region, the complete plastid genome, and transcriptomes, have resulted in numerous polytomies and clades that contradicted the morphology-based classification for several sections (Vijayan et al., 2009; Jiang et al., 2010; Yang et al., 2013; Huang et al., 2014; Wu et al., 2022; Zan et al., 2023). In Theeae, intergeneric relationships, especially the position of *Apterosperma* and *Laplacea*, have been challenging to resolve. The two most recent study, using either 610 low-copy nuclear genes or 1785 nuclear genes from transcriptomes, reported contradict relationships in Theeae with medium support values (Cheng et al., 2022; Zhang et al., 2022).

The rapidity of the speciation itself is likely to be responsible for the difficulty in resolving the phylogenetic history of Theeae and *Camellia*. Several instances of introgression and hybridizations have been hypothesized and identified within *Camellia*, within *Stewartia*, and within Gordonieae, in studies using RAD-seq and transcriptome data (Lin et al., 2019; Zan et al., 2023; Zhang et al., 2022). ILS, another source of discordance between gene trees, has not been reported in Theaceae so far. There is still lack a quantitative analysis to evaluate the discordances among gene trees and species tree, and processes underlying the discordances, across deep relationships and major clades of the Theaceae family, partly due to the computational and method limitations.

To unravel the phylogeny of this complex family and quantify the impact of possible biological processes during evolution, we sampled multiple low-copy nuclear genes included in the recent universal probe set Angiosperm 353 (Johnson et al., 2019) generated by the target enrichment method from herbarium specimens and inferred phylogenies for 68 representatives of Theaceae species. The target enrichment technique enables researchers to recover large amounts of nuclear sequences for divergent non-model species (Cronn et al., 2012; Grover et al., 2012; Straub et al., 2012; Schmickl et al., 2014) and can be applied to highly degraded samples, making it valuable for analyzing historical specimens or species challenging to identify (Bieker and Martin, 2018; Brewer et al., 2019). The probe set we used has been employed effectively across various taxa to investigate intergenic and interspecific relationships, even for groups that have experienced rapid speciation (Baker et al., 2022, 2021). In addition to resolve the intergenic and intragenic relationships of Theaceae, we also quantitatively evaluated discordances of phylogenies and quantified the relative contributions of ILS, introgression, and gene tree estimation errors to the observed gene tree discordances.

2. Materials and methods

2.1. Taxon sampling

We sampled 76 herbarium specimens spanning from 1917 to 2018. To ensure that species identifications were reliable, we aimed to select specimens that were collected from within the core distribution range of the species and by expert collectors such as Shui Ying Hu, Heng Li, and B. Bartholomew. Specifically, the samples include 44 species of *Camellia*, eight species of *Schima*, six species of *Stewartia*, five species of *Polyspora*, two species of *Laplacea*, five species of *Pyrenaria* and the sole species in *Franklinia*. The species in *Camellia* spanned 10 out of the 14 sections followed Ming (1999) and 14 out of the 22 sections (64 %) as per Chang's, 1998 classification. We aimed to have a nuclear dataset with coverage comparable to the currently most complete chloroplast genome dataset (Yan et al., 2021). For species that have many cultivated varieties, we sampled two specimens from different regions to reduce the bias of potential hybridization. We selected two species in Pentaphylacaceae as outgroups (*Andinandra millettii* and *Pentaphyllax euryoides*). The information on voucher specimens, the standardized species names according to the latest World Flora Online database (WFO, 2021), and taxonomical information from the Flora of China (Min and Bartholomew, 2007) are listed in Table S1.

2.2. Target enrichment probe set

We used the recently published target enrichment probe set "Angiosperm353″ to capture multiple nuclear genes from the samples. This probe set is designed to capture 353 single-copy nuclear genes from any flowering plant, making it feasible to resolve both shallow and deep phylogenetic relationships, a prerequisite for resolving the problematic regions of the Theaceae phylogeny. A previous study showed that this probe set could recover ~ 150 nuclear genes on average for species within Ericales (the order that Theaceae belongs to), with total lengths recovered for both coding and non-coding regions around ~ 250kbp on average (Johnson et al., 2019). This suggested that the probe set could be effectively used on Theaceae.

2.3. DNA isolation, library preparation, sequencing, and data assembly

For each DNA extraction, approximately 15 mg of leaf tissue was used in a modified CTAB protocol. The extracted DNA was quantified using a Qubit 3.0 fluorometer (Life Technologies, Carlsbad, California, USA) and 4200 TapeStation System (Agilent Technologies, Santa Clara, California, USA). We prepared the dual-indexed libraries with 2–48 ng input genomic DNA using the Kapa DNA Hyper Plus Library Prep Kit (Kapa Biosystems) at 1/4 the recommended volume and size selected for 250–700 bp. The final amplified library was combined at equal ratios resulting in 10–20 indexed samples and a total of > 100 ng libraries per tube. We followed the protocol of the Angiosperm353 baits kit and

enriched low-copy nuclear genes. Hybridization was carried out at 65 °C for 28 h. After target enrichment, the subsequent libraries were pooled together and sequenced on Illumina Nextseq for 150 bp paired-end reads, and Illumina Hiseq for 125 bp paired-end reads at the Bauer Core Facility of Harvard Faculty of Arts and Sciences.

We removed the adapter and cleaned the raw reads using *fastp* with default settings (Chen et al., 2018), and assessed the quality before and after cleaning using FastQC (Andrews, 2010). The qualified reads were processed into the HybPiper pipeline (Johnson et al., 2016) with default settings, to assemble the targeted sequences using the target file (available at https://github.com/mossmatters/Angiosperms353). Additionally, we extracted supercontigs and detected possible paralogs in the dataset using the scripts in the Hybpiper pipeline. We calculated the target enrichment efficiency, recovered length, and the coverage of the recovered sequences using R 3.5.3 (R Core Team, 2019). All the assembly was done on the Odyssey cluster of Harvard University.

2.4. Alignment and data filtering

The multi-fasta files generated by Hybpiper for each locus were aligned individually using the "auto" setting in MAFFT v.7.407 (Katoh and Standley, 2013). The alignments were then trimmed in several steps. First, all sites occurring in at least 50 % of the samples (referred to as "good positions") were identified. Then "resoverlap 0.5" was used to exclude all sites that were not good positions and "seqoverlap 0.5" was used to trim away sequences with less than 50 % good positions. Finally, poorly aligned regions were trimmed away using the "automated1" setting in trimAl v.1.2 (Capella-Gutiérrez et al., 2009). The trimmed alignments were proceeded to RAxML v8 (Stamatakis, 2014) to estimate the best maximum likelihood tree under the GTRGAMMA substitution model with 100 bootstrap trees. To eliminate the influence of long-branch attraction (LBA) (Philippe et al., 2011), we ran TreeShrink (Mai and Mirarab, 2018) on the best tree and the corresponding sequences for each locus. Sequences corresponding to long outlier branches (detected with false positive tolerance (α) set to 0.05) were removed from the alignment. We then reran RAxML v8 (Stamatakis, 2014) with the same settings for each of the filtered alignments. The resulting gene trees were used to estimate the species tree with the summary coalescent method. The filtered alignments of each locus were then concatenated and used for a supermatrix analysis (see Fig. S2 for a flowchart summarizing the process). Summary statistics for gene trees, i.e., the average bootstrap value and average branch length, were calculated in R using package ape (Paradis and Schliep, 2019). Summary statistics of alignments were calculated using AMAS (Borowiec, 2016).

As the enrichment efficiency differed across different groups (Fig. S1), the resulting sequences had issues deriving from unbalanced sampling and phylogenetic noise. We addressed the sensitivity of our results to these issues by investigating the effect of applying various filtering strategies to the full trimmed data (referred to as the L0 dataset), including locus filtering, site filtering, and taxon filtering. Filtering on loci was done by either (a) removing loci with high levels (>20 %) of missing data (the L2 dataset) or by (b) removing loci with a low level of informativeness and high evolutionary rate (based on a "locus informativeness score" – see Supplementary Note S1; the F3 dataset). Filtering on sites was done by removing sites with particularly high (>5) substitution rates (the PI5 dataset). Filtering on taxa was done by removing samples from which few (<60 %) loci were recovered before the trimming step in the above paragraph (the T02 dataset). We investigated the phylogenetic informativeness of datasets generated after different filtering approaches, based on the number of parsimony sites and the average bootstrap value for each gene tree, and compared the resulting phylogenies. More details about the filtering strategies are given in Supplementary Note S1 and Table S2.

2.5. Phylogenetic analysis

We inferred the species tree using two different methods, a concatenation method that estimates the species tree based on a concatenated dataset that assumes all genes share the same evolutionary history (though partitions of the alignment may have different substitution rates), and a coalescent method that estimates the species tree based on combining individual gene trees, which allows loci to have different evolutionary histories and performs better than concatenation method under high level of ILS (Yang, 2006).

For the concatenation approach, we first obtained the best partition scheme and the best rate model (within the GTR model family) for each partition of the supermatrix using PartitionFinder (Lanfear et al., 2016), allowing for evolutionary rate heterogeneity among partitions. Then we searched for the best maximum likelihood tree with RAxML-ng (Kozlov et al., 2019) using 10 random and 10 parsimony starting trees to ensure a comprehensive search of tree space. We assessed node support by generating 1000 non-parametric bootstrap replicates. The bootstrap values were then mapped to the best tree.

For the coalescence approach, we inferred species trees using the summary quartet-based method in ASTRAL-III (Zhang et al., 2018). We tested the influence of gene tree estimation errors by collapsing to polytomies all nodes with either < 10 % (given code BS10), <30 % (given code BS30), or < 50 % (given code BS50) bootstrap support values and compared the resulting trees to the tree without collapsed nodes. This procedure was applied to each best fitting ML locus tree using Newick utilities (Junier and Zdobnov, 2010), before feeding the contracted trees to ASTRAL-III (Zhang et al., 2018). The support for individual nodes was calculated from the default posterior values.

We selected L0-BS10 tree (i.e., the tree derived from applying the coalescent approach to the unfiltered dataset while collapsing nodes with less than 10 % bootstrap support) as the "best" species tree under coalescent method (Table S4). The tree has the highest average posterior support value compared with other phylogenies based on loci filtered datasets (F3 and L2). We identified F3 tree as the "best" species tree under concatenation method. The tree has the highest average bootstrap support value among phylogenies based on loci-filtered datasets (Table S3).

2.6. Organelle and ribosomal phylogeny

We expanded an existing dataset of plastome and ribosomal DNA from Yan et al. (2021) with newly assembled data sequenced for this study. The assembly process of plastome and ribosomal DNA followed Yan et al. (2021). The combined dataset included 63 samples, and 30 of them were from the same specimens used in this study. The phylogeny was built using partitioned maximum likelihood in RAxML-ng with 1000 bootstrap replicates.

2.7. Discordance evaluation

The combination of different phylogenetic inference methods, filtering approaches and multiple genes resulted in multiple alternative phylogenies. The discordance among phylogenies was evaluated two-fold. First, we compared all species trees across the two phylogenetic inference methods and the different filtering approaches. Second, for the coalescence approach, we calculated the variation between gene trees and the species tree.

We quantified the stability of nodes to filtering methods using the gene concordance factor (gCF), calculated as the frequency with which a branch from the "best" species tree (L0-BS10) appeared in the trees estimated using the filtered datasets. The measurement complements the traditional bootstrap values and posterior support values, allowing us to see if the topology uncovered in the "best" species tree is a dominant tree topology among the species trees (indicated by high values of the concordance factor) and insensitive to filtering methods.

The factor was calculated by using the option "-gcf" in IQ-TREE v 1.7 (Nguyen et al., 2015).

We quantified dissimilarity among trees by the normalized quartet score (norQS) with the R package Quartet (Sand et al., 2014; Smith, 2019). This score is calculated as:

$$norQS = \frac{d1 + d2}{d1 + d2 + 2s}$$

where d1and d2 is the number of quartets resolved differently in each tree, and s is the number of quartets resolved in the same manner in the two trees (Kuhner and Yamato, 2015). The norQS ranges from 0 (same topology) to 1 (completely different topology). The distance matrices were then scaled using principal coordinate analysis to a two-dimensional space for easy visualization (Hillis et al., 2005). We evaluated the influence of exon characteristics on each gene tree by running linear models for key exon characteristics and the norQS of the gene tree vs. the best species tree obtained using coalescent and concatenation methods, respectively. These exon characteristics include the number of parsimony informative sites, the missing data percentage, and the average bootstrap value.

To evaluate the consistency with which the generated sequences support the final topologies, we calculated two measures of genealogical concordance: gCF as mentioned above, calculated as the frequency with which a branch from the species tree appeared in the individual gene trees; and the site concordance factor (sCF), which is defined as the percentage of decisive sites in alignments supporting a branch in the species tree. sCF was calculated by using the options "-scf" in IQ-TREE v 1.7 (Nguyen et al., 2015). Low values of concordance factors imply a conflict among gene trees that might stem from an alternative dominant tree topology or from many low-frequency alternative gene topologies (Villaverde et al., 2018). Two measures of decisiveness, the number of decisive genes (gN, number of gene trees that contained the terminals of a branch) and the number of decisive characters (sN) were also calculated in IQ-TREE (Nguyen et al., 2015) and plotted against branch lengths estimated in RAxML.

2.8. Quantifying ILS, reticulation events, and gene tree estimation errors

We investigated the relative contributions of incomplete lineage sorting (ILS), gene flow (e.g., introgression), and gene tree estimation errors in driving gene tree – species tree discordance, following a recently published variance decomposition method (Cai et al., 2021).

The gene tree variation of each node in the "best" species tree was quantified by the quartet support values for all the best-scoring maximum likelihood gene topologies (collapsing nodes with less than 10 % bootstrap support) generated by RAxML for each locus. Gene flow and ILS were represented by the reticulation index and the theta value, respectively, as calculated following Cai et al. (2021). The branch length in coalescent units, used to estimate theta values, were calculated in ASTRAL III. For gene tree estimation errors, we simulated 348 gene alignments from the "best" coalescent-based species tree using Seq-Gen (Rambaut and Grassly, 1997). The alignment size of each gene was set to 400 bp, the mean size of the empirical dataset. Each alignment had unique substitution model parameters estimated from the corresponding empirical alignment by RAxML. We generated phylogenies for each of these simulated alignments using RAxML, collapsed nodes with bootstrap value less than 10 %, and summarized the quartet support for each node on the species tree using ASTRAL -t 1. This approach helped eliminate the bias brought by missing data in the alignments and low-support nodes in gene trees. Higher quartet support values here imply that a smaller proportion of gene tree variation may be attributed to gene tree estimation error. Finally, we partitioned the relative contribution of each factor using linear regression in the R package relaimpo (Groemping and Matthias, 2021). We log-transformed theta values during partition due to the high skewness of the variable.

To address the possibility of reticulate relationships within *Polyspora*, phylogenetic network analyses for this genus were also carried out using PhyloNet 3.8.0 (Than et al., 2008; Wen et al., 2018). More details about the method and results are given in Supplementary Note S1, Table S5, and Figure S8).

3. Results

3.1. Characteristics of the datasets and the efficiency of target capture sequencing

We succeeded in getting sequence data from all 76 specimens. The average percentage of reads mapped to the target genes was 28.7 % (12.5 % − 41.8 %; SD 0.065). We recovered a total of 348 exons of the 353 targeted loci with more than three samples at a depth of coverage > 8x. After excluding 5 samples with an extremely low number of loci (<5) or enrichment efficiency (<20000 bp), the remaining 70 samples had an average number of 287 loci with contigs, among which 138 had a length > 50 % of the target genes, and 69 had length > 75 %. For taxon recovery, each locus captured 56 samples on average. Of the recovered loci, 69.1 % (242 loci) were represented in more than 80 % of the samples (56 samples).

The length of alignments for each locus before trimming was 108–2761 bp (mean = 737 bp) with an average missing data percentage of 46 % (0.4 % − 77.5 %) (Supplementary Fig. S1). After removal of misaligned bases, gappy regions, and short sequences in each alignment, loci were 59——1902 bp long (mean = 400 bp), and each with 1——471 parsimony-informative characters (mean = 61). The final baseline combined dataset (exons_L0) was 139540 bp long, including 18,909 (13.6 %) parsimony-informative characters, 40,227 variable sites, 40.6 % missing, and a GC content of 45.6 % (Table S2). A total of 16 exons were identified with potential paralogs; we selected the contigs that were more likely to be homolog based on gene trees. The contig combined dataset (contigs_L0) was 145089 bp long after trimming, including 25,581 (17.6 %) parsimony informative characters, 59,099 variable sites, and 39.3 % missing.

3.2. Phylogenetic relationships among and within genera

Overall, the phylogenies generated from the nuclear enrichment dataset agreed with the updated plastome-based phylogeny on the following relationships: (1) the monophyly of the three major tribes and their relationship as Stewartiea+(Gordoniea + Theeae) (bootstrap support (BS) = 100 %/ posterior probability (PP) = 1.00 in the nuclear dataset); (2) the generic monophyly of *Stewartia*, *Schima* and *Polyspora* species; (3) the placement of sect. *Calpandria* (*Camellia lanceolata*) as the basal lineage of *Camellia*.

The intergeneric relationships were generally consistent between the concatenation analysis and the coalescence analysis, as well as with the plastome-based phylogeny. The only exception was the placement of the genus *Laplacea*, which the coalescent-based species tree recovered as sister to the other genera in Theeae (PP = 0.88 in the L0 dataset), whereas the concatenation method recovered it as sister to *Pyrenaria*. The plastid tree recovered it as sister to *Polyspora* and *Camellia* (BS = 69 %) (Fig. 1, Supplementary Fig. S4).

Within genera, we found some discordances between our new nuclear phylogeny and the plastome phylogeny (Fig. 1). Within *Camellia*, the relationships recovered by the nuclear phylogeny, however, are consistent with the most recent morphology-based classification system of Min and Bartholomew (2007) to some extent. We recovered four of their sections forming two major clades with strong support: One is *Theopsis* + *Camelliopsis;* the other is *Camellia* + *Paracamellia*. The nuclear phylogeny also recovered the sections *Tuberculatae* and *Thea*. The basal position of section *Calpandria* in *Camellia* was confirmed by both the nuclear and plastome datasets. However, a few discrepancies with the morphology-based classification remained: The seven species of section

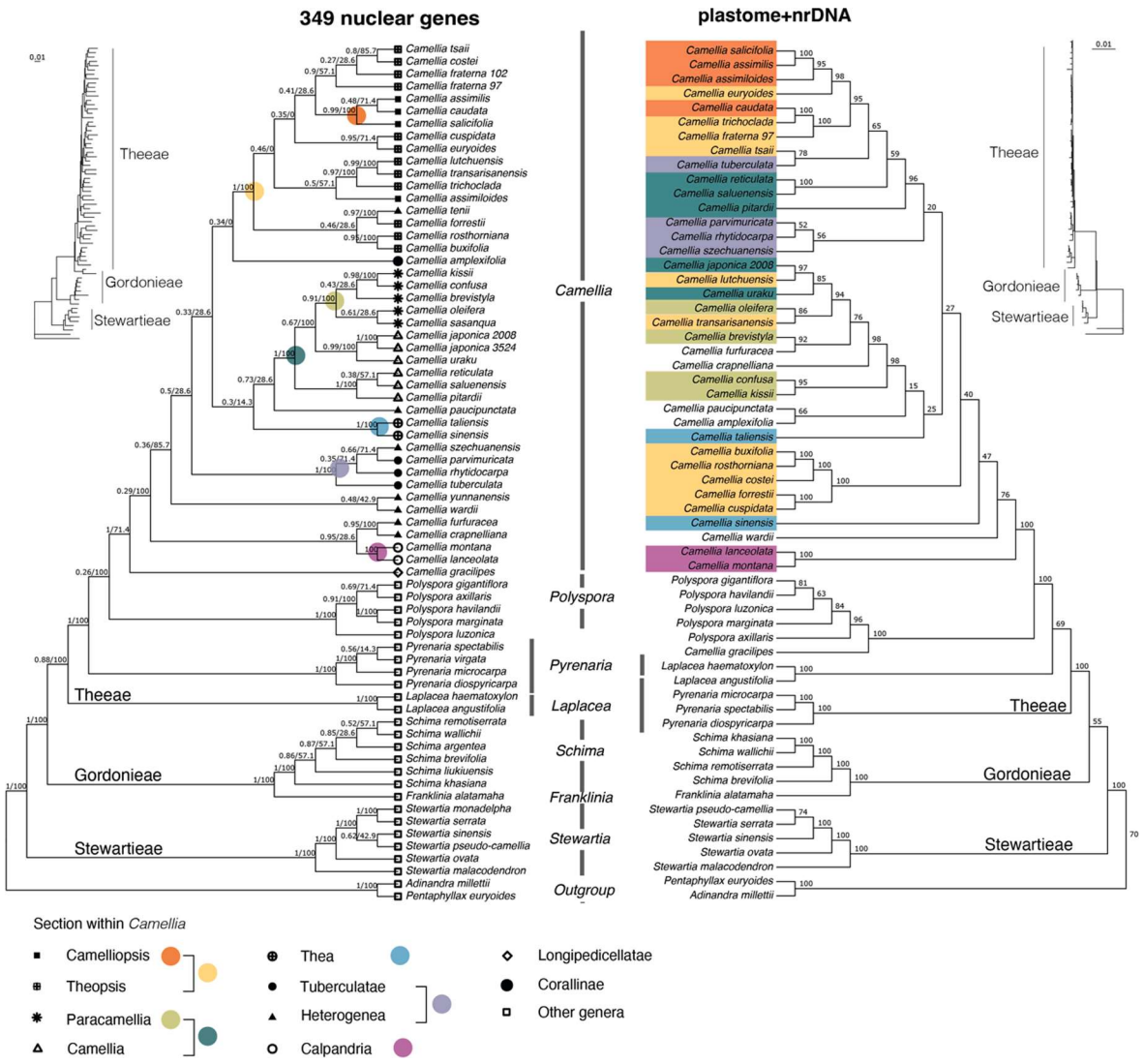


Fig. 1. Phylogenetic relationships in Theaceae inferred by the coalescent method in ASTRAL III based on the nuclear enrichment dataset L0-BS10 (left) and that inferred by the concatenation method in RAxML based on the plastome + ribosomal dataset (right). The branch annotations on the tree on the left show the posterior support and the concordance factor across all different filtering approaches. Nodes marked with a colored dot are highly supported clades referenced in the paper and correspond to the colored species on the right. The tips are annotated with different shapes representing sections in *Camellia*. The nodes on tree on the right display the bootstrap values. The plots in the upper left and upper right represent the phylogram of the same trees showing branch lengths.

Heterogenea were scattered throughout the phylogeny with low to moderate support values, indicating that this group may be paraphyletic (Fig. 1). Finally, *Camellia gracilipes* (the only species in our dataset from sect. *Longipedicellatae*) was clustered with *Polyspora* (BS = 100 %) in the plastome dataset, while it was grouped with all the other *Camellia* species (PP = 1.00) in the nuclear dataset.

3.3. Stability of nodes to filtering

Under the coalescent method, the relationships among genera were generally stable to different types of filtering, also for those genera that were only supported with low posterior probabilities. The phylogenies generated after collapsing weakly supported nodes to polytomies (the BS10 trees) were particularly stable to filtering (mean quartet dissimilarity from 0.09 to 0.10) (note the small minimum spanning ellipses on Fig. 2A). The only exception to the stability of genera was again *Camellia gracilipes*, which under some filtering approaches got moved into *Polyspora*. Because of the stability and high support value, we used the best species tree under the coalescent method (L0-BS10) as the overall "best" species tree for the dataset.

Within genera, most relationships between deeper nodes in *Stewartia*, *Polyspora*, and *Pyrenaria* were stable to filtering (gCF = 100, Fig. 1). In *Camellia*, in contrast, the nodes that were most stable to filtering were mostly shallow. Generally, the relationships within *Schima* were quite sensitive to filtering, as were the relationships among the deeper nodes of *Camellia* where branch lengths were short (Fig. 1). Excluding loci with a high percentage of missing data (L2) or low informativeness (F3) slightly reduced the average node posterior probability, because it decreased the number of highly supported nodes. Filtering out fast-evolving sites (PI5) slightly increased the average posterior probability of nodes (Supplementary Fig. S6). Collapsing low-support branches (BS10 and BS50) did not always increase the mean posterior probabilities but significantly increased the normalized quartet scores (Supplementary Table S4).

The phylogenies generated by the concatenation approach were more sensitive to filtering than those generated by the coalescent approach (mean quartet dissimilarity = 0.11 for filtered and unfiltered datasets) (Fig. 2A). The tree with the highest bootstrap support (an average of 80.3 %) was obtained by removing fast-evolving positions (PI5). Filtering out loci with a high percentage of missing data (L2) did

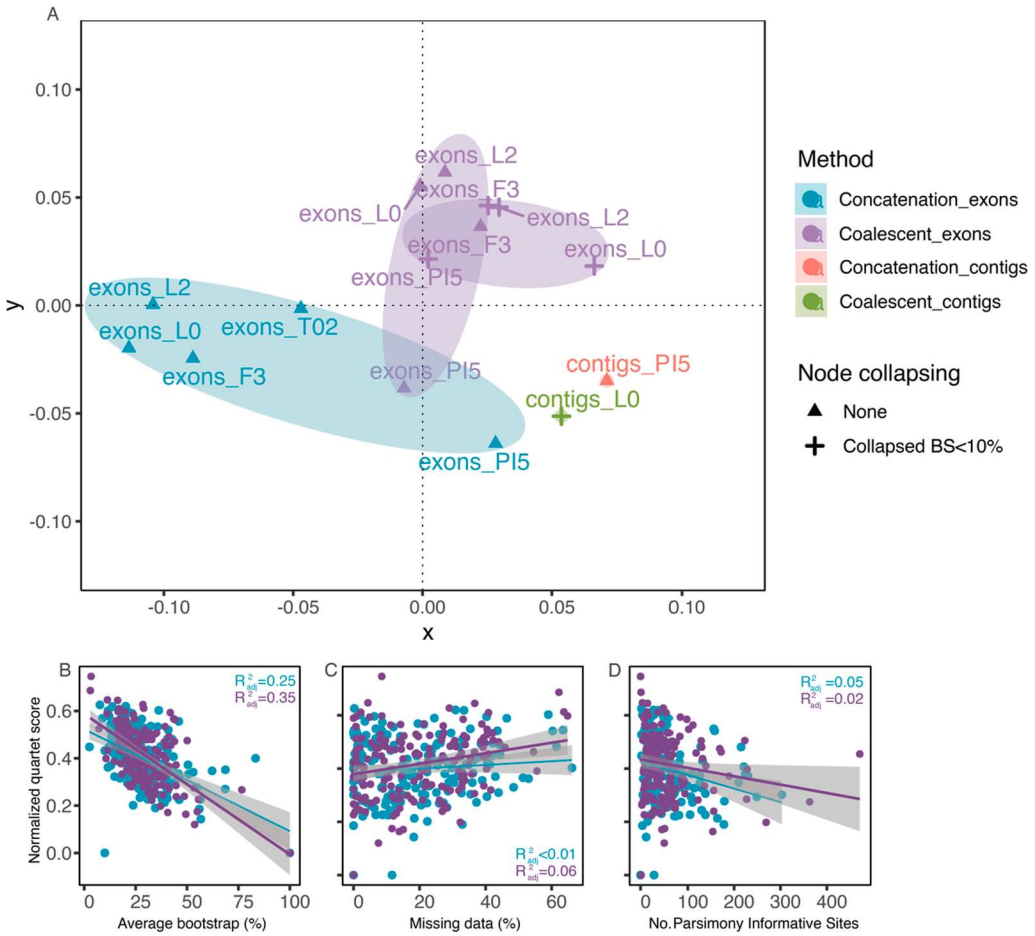


Fig. 2. A. Quartet dissimilarity between species trees inferred by coalescence or concatenation approaches and filtered according to different criteria. The codes are defined in Note S1 and Table S2. The dots are surrounded by minimum spanning ellipses for each combination of phylogenetic approach and polytomy criterion. B-D. Scatter plots of quartet dissimilarity between exon gene trees and the species tree against key gene tree characteristics, with OLS regression line and 95 % confidence band. The corrected goodness-of-fit of the regression is annotated on the plots. Purple color represents the best coalescent tree (exons_L0-BS10) and green color represents the best concatenation tree (exons_F3). B. Average bootstrap value of gene trees. C. Missing data percentage. D. Number of parsimony sites. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

not have an impact on the overall bootstrap value, while excluding loci with low phylogenetic informativeness (F3) increased support for branches that were poorly supported by the full (L0) dataset (Supplementary Table S3). Filtering samples with a moderate level of missing data (T02, based on 53 samples) increased node support values somewhat under both concatenation and coalescent methods (Supplementary Table S3). The alignment length in the T02 data was 9047 bp longer than the L0 dataset, with less missing data and more parsimony informative sites (Supplementary Table S2), but overall recovered the same topology. The supercontig datasets also resulted in the same topology as the L0 dataset, although with higher support values (Fig. 2A, Fig. S5).

3.4. Incongruence evaluation between gene trees and species trees

Individual gene trees exhibited considerable topological disagreement with the overall species tree. The average quartet dissimilarity for the full taxon sets (L0, F3, PI5, L2) was 0.41, showing that on average only 59 % of the quartets found in the common tip set were identical. The average quartet dissimilarity of the reduced sample dataset (T02) was a little lower at 0.37. (Supplementary Table S4). The exon datasets showed a relatively strong negative linear correlation between the dissimilarity of gene trees from the species tree and the average bootstrap value of gene trees (Fig. 2B). This relationship was stronger for the coalescent species tree (adj. R^2 = 0.35) than for the concatenation species tree (adj. R^2 = 0.25). There was no significant relationship between the gene trees – species tree dissimilarity and the amount of missing data, nor with the number of parsimony informative sites (Fig. 2C-D).

For internal branches, both gene tree concordance factors (CF) were positively correlated with branch length (Supplementary Fig. S6), i.e., shorter branches tended to have lower CF values. The gCF values were generally lower than sCF, especially for nodes within *Camellia*, a common observation when gCF values are affected by stochastic error. Many nodes with a bootstrap value of 100 % in the RAxML trees had low gCF and sCF values (Fig. 3A), another frequently observed phenomenon revealing that even though nodes consistently appear in sample trees, there is considerable variation in the relationships encoded by the underlying sequence data. The gCF was high for the basal branch of each of the three tribes (Gordonieae gCF = 78.8 %, Stewartieae gCF = 69.8 %, Theeae gCF = 63.3 %), and it was higher than 50 % for *Laplacea* (gCF = 65.9 %), *Schima* (gCF = 51.8 %) and the sister relationship of *Stewartia serrata* and *S. monadelpha* (gCF = 52.9 %). The sister relationship of the Stewartieae to the Gordonieae + Theeae clade had a gCF value of 40.2 %. The rest of the intergeneric relationships within Theeae and interspecific relationships had much lower gCF, especially around the backbone of *Camellia*, where several relationships were only supported by one or a few gene trees (Fig. 3B). However, the number of decisive sites for many nodes with low gCF values was not significantly lower than that of nodes with high gCF values, showing that the informative sites are possibly scattered in different genes. The sCF value also showed

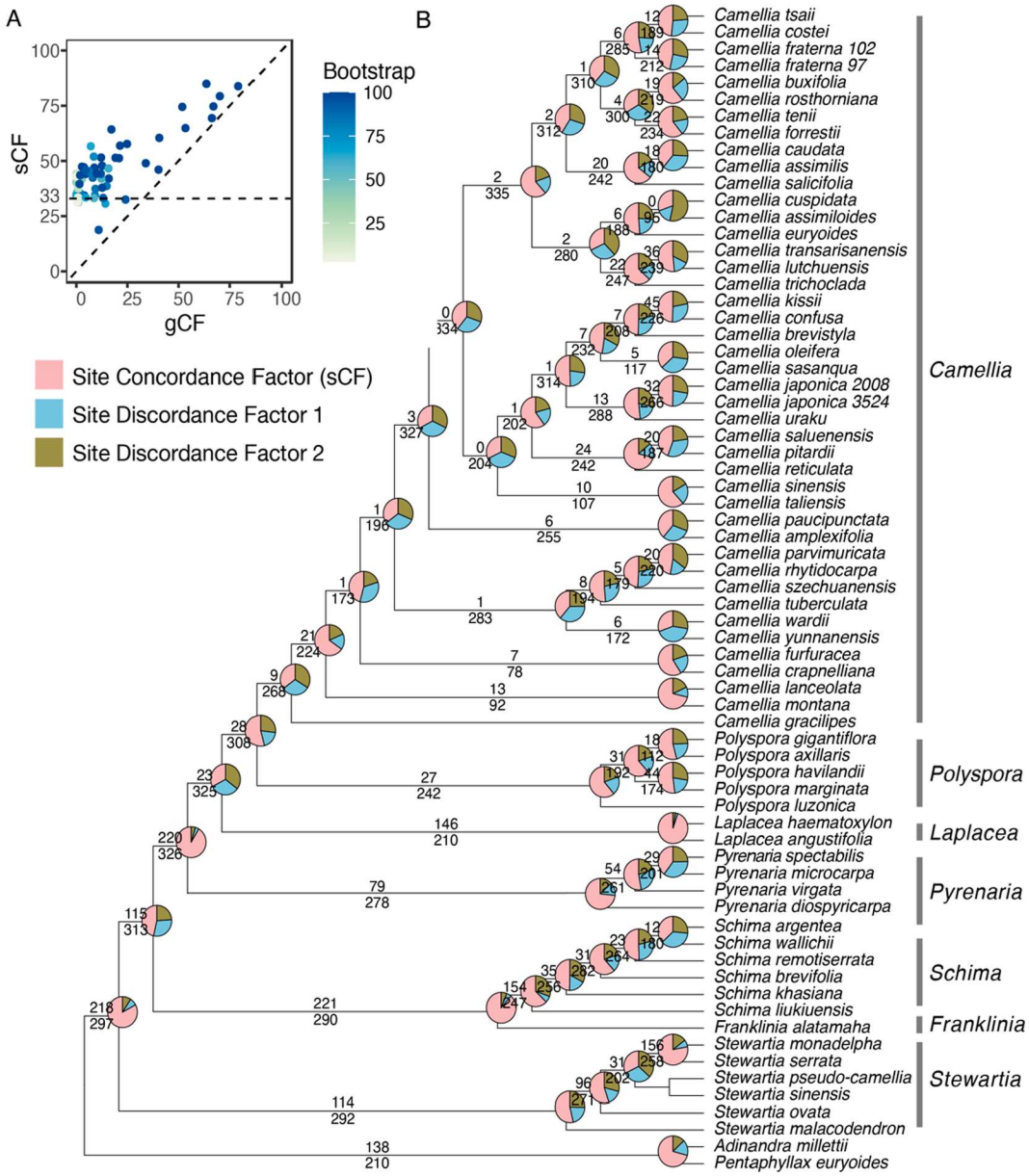


Fig. 3. A. The relationships between gCF, sCF, and RAxML bootstrap values of nodes from the L0 dataset (i.e., without filtering). The vertical dashed line indicates an sCF value of 33, the expected minimum implying a lack of information in the data. B. Patterns of site concordance and conflict based on the concordance analysis. The pie charts at each node show the proportion of sites in concordance (pink) and discordance (blue and brown). The numbers above and below each branch are the numbers of concordant and total genes at each bipartition, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

that there is no overwhelming support for any particular resolution for those nodes (Fig. 3B).

Regarding the impact of the filtering strategies on CF values, we observed that removing fast-evolving sites (PI5) increased sCF values substantially, especially for short branches, indicating that fast-evolving sites had a substantial impact on the variation among inferred topologies. All filtering approaches on loci or sites increased the gCF values to some extent, especially for longer branches (Supplementary Fig. S6).

3.5. Distribution and contribution of different factors to gene tree discordances

Our investigations revealed positive correlations between the degree of ILS, gene tree estimation errors, and actual gene tree variations (Fig. 4E-G). ILS, gene flow, and estimation errors in total explained 63 % of the observed variations in gene-tree heterogeneity. Within the explained variation, gene tree estimation error explained most (77 %), followed by ILS (22 %) and gene flow (0.3 %). We observed different distribution patterns of the three factors across different clades. The rapidly evolving clade along the backbone of *Camellia* (Fig. 4D) was also where we observed the greatest variation among gene trees, relatively low simulated gene-tree support, and the highest theta values, indicating the likelihood of extensive ILS and gene-tree estimation errors (Fig. 4B-C). Low simulated gene-tree support was found within *Schima* with low introgression and ILS level, which indicated that estimation errors might be the main reason for the gene-tree heterogeneity in this clade. The reticulation index indicated that introgression might have happened in the *Polyspora*, *Stewartia* and in a few nodes in *Camellia* and contributed to the gene-tree variations there (Fig. 4A).

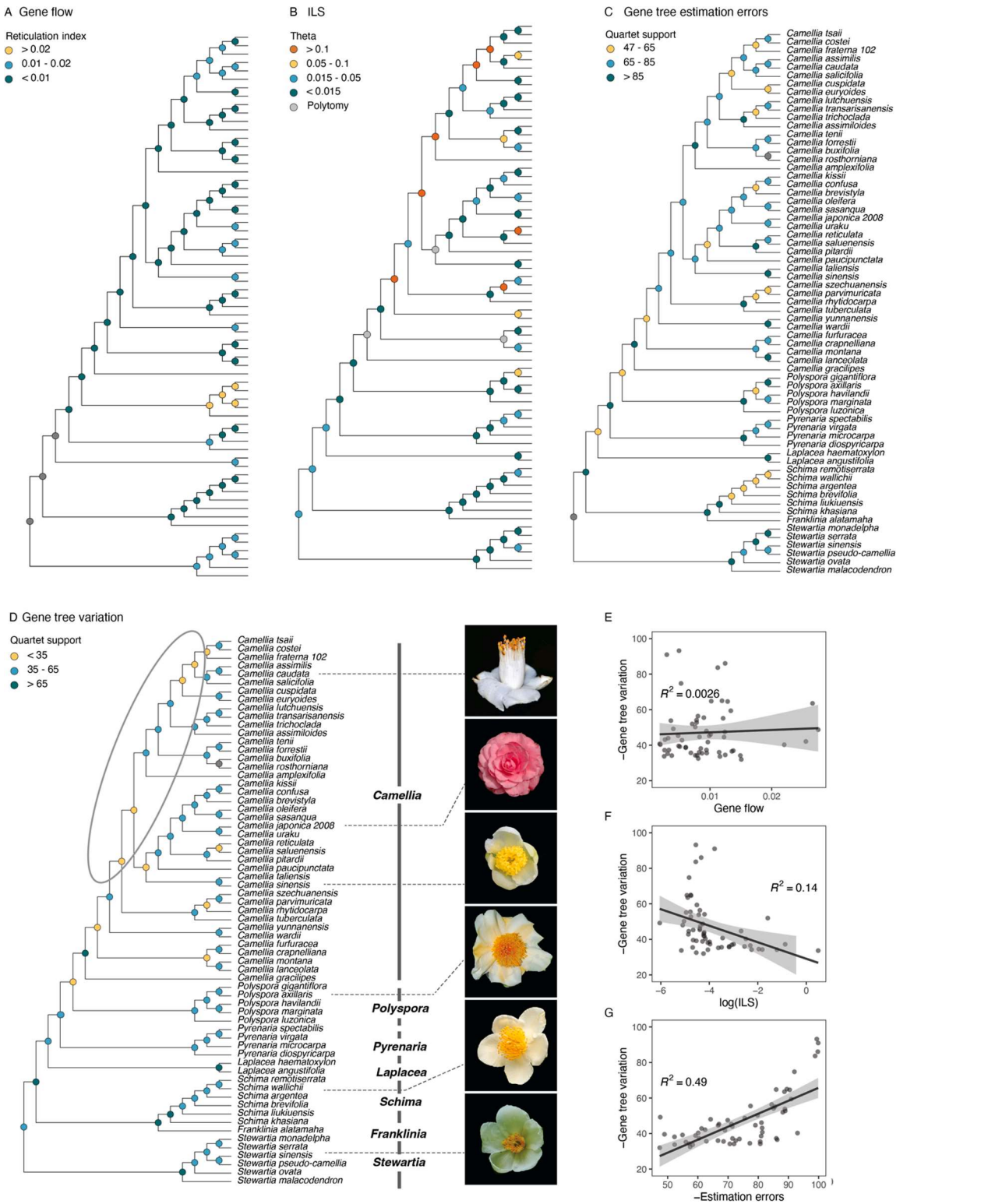


Fig. 4. Phylogenetic distribution of gene flow, ILS, gene tree estimation error, and gene tree variation across Theaceae. A. Gene flow. Nodes are colored by Reticulation Index. Warmer colors indicate a higher percentage of unbalanced gene trees and thus a higher level of gene flow. B. ILS. Nodes are colored by the inferred population mutation parameter theta. C. Gene tree estimation error. Nodes are colored by simulated quartet support. Lower values represent a higher level of gene tree estimation error. D. Gene tree variation. Nodes are colored by empirical quartet support. Photos on the right from top to bottom show the flowers of the main groups: *Camellia caudata, C. japonica, C. sinensis, Polyspora axillaris, Schima wallichii, Stewartia sinensis* (from ppbc.iplant.cn). E-G. Relationships between gene flow, ILS, gene tree estimation errors, and gene tree variations.

4. Discussion

Our ~ 350 target enriched nuclear gene dataset resolved most of the intergeneric relationships in Theaceae. The topology was broadly consistent with the plastome-based phylogeny and with other recently published phylogenies based on a range of different approaches. As such, most of the overall relationships within the family appear to be generally resolved.

Our analysis also addressed historically difficult relationships along the rapidly radiating backbone of tribe Theeae and within the genus *Camellia*. Our multi-locus nuclear coalescent-based phylogeny reestablished the sections from the most recent morphology-based taxonomy of Ming (1999) and Min and Bartholomew (2007), which has otherwise been questioned by several recently published molecular analyses. However, the relationships between sections remain weakly supported, and further analyses reveal a high degree of gene tree heterogeneity and low concordance factors, which could broadly be attributed to gene tree estimation errors (with a relatively smaller contribution of incomplete lineage sorting). Applying a range of filtering approaches did not qualitatively change the stability of the topology, and it thus seems unlikely that the most complex relationships within e.g., *Camellia* can be resolved, except possibly with more focused sequencing of greater numbers of long and informative nuclear genes with introns.

4.1. A new multi-locus consensus nuclear phylogeny for Theaceae

The relationships between tribes were resolved in agreement with those resolved using plastid genomes and transcriptomic datasets (Yan et al., 2021; Cheng et al., 2022; Zhang et al., 2022). The same is true for the intergeneric and most interspecific relationships resolved for Stewartieae (*Stewartia*) and Gordonieae (Lin et al., 2019; Cheng et al., 2022; Zhang et al., 2022).

Within *Camellia*, our analysis recovered relationships among most major clades in the morphology-based classification system by Ming (1999), though with some minor adjustments. The 15 sampled species from section *Camelliopsis* (or *Eriandra* in Chang, 1998) and sect. *Theopsis* were intermixed, confirming the monophyly of a clade constituting both sections, as suggested by (Vijayan et al., 2009). We recovered part of the backbone as *Thea*+(*Camellia* + *Paracamellia*), which is sister to *Camelliopsis* + *Theopsis* in the coalescent phylogeny. Although the support values were low, the relationships align with the phylogenies constructed using whole-transcriptome sequencing data (Fig. 1) (Xia et al., 2017; Zan et al., 2023) and we reaffirmed that the designation of sections was quite different from the morphology scheme by Chang (1998). The major clades were slightly different from a recent taxonomically comprehensive study of genus *Camellia* using three nuclear markers (Zhao et al., 2023). However, we also found that *Camellia japonica* did not group with other sect. *Camellia* species but grouped with sect. *Paracamellia* instead. The two specimens of *C. lanceolata*, one from Indonesia and *C. montana* (a synonym of *C. lanceolata*) from the Philippines, grouped together with high support in both nuclear and the plastome phylogenies. They formed a clade with two species from Chang's sect. *Furfuracea*, namely *C. furfuacea* and *C. crapnelliana*, suggesting the close relationships between these species (Zhao et al., 2023).

A notable result is the polyphyly of Ming's section *Heterogenea* Sealy. In our study, the positions of the seven sampled species were not stable across different datasets, though they never formed a single clade. These species were classified into five different sections according to Chang, 1998. It is thus highly doubtful whether Ming's treatment of *Heterogenea* should be accepted, a sentiment echoed by other taxonomic studies based on leaf structures and nuclear markers (Jiang et al., 2010; Zhao et al., 2023). Among the seven *Heterogenea* species, *Camellia tenii* grouped with three other *Theopsis* species, *Camellia forrestii, Camellia rosthorniana, and Camellia buxifolia,* with high support values in almost all phylogenies. Interestingly, the species is classified under the *Paracamellia* section according to Chang's system. Such unstable placement of *Camellia tenii* indicates a necessity to reevaluate the taxonomic placement. While our study is the first to include this species in a phylogenetic analysis, more specimens and individuals should be sampled and analyzed before drawing a solid conclusion.

Camellia gracilipes Sealy, classified in Ming's section *Longipedicellata*, clustered with *Polyspora* in our updated plastome + ribosomal tree with strong support. However, it was placed basal to *Camellia* in most coalescence-based and concatenated-based nuclear phylogenies with high support value (Fig. 1), similar to Zhao et al., (2023). Given the consensus among most nuclear gene trees, this conflict might result from chloroplast capture (Fig. 4D). Incorrect identification of the species is unlikely. The specimen of *C. gracilipes* we sampled was collected from Vietnam and was determined by Theaceae specialist Joseph Robert Sealy. The fruit of the specimen, a typical *Camellia* capsule fruit with one seed, was also compared with the type specimen (Supplementary Fig. S7). The species has a relatively wide distribution from Vietnam to South China, overlapping with that of *Polyspora*, which provides conditions for potential introgression (Min and Bartholomew, 2007).

Apart from the relationships discussed above, the deeper relationships among clades within *Camellia* were not well-resolved. A more comprehensive sampling of species within these questionable clades from different regions, especially the less-studied Southeast Asia, are needed in the future study to refine the above taxonomic treatment. Our results also show the importance of sequencing type specimens and herbarium specimens identified by clade experts in clarifying taxonomy.

4.2. Addressing gene-tree heterogeneity in rapidly evolving clades

We observed considerable incongruence between phylogenies constructed using different datasets and methods (Figs. 1 and 2). As described above, our nuclear coalescent tree identified sub-clades within *Camellia* that were more closely aligned with the classic morphological classification compared to those identified by the plastome-based tree and the nuclear concatenation tree. It is possible that biological processes during evolution have complicated the identification of distinct clades within this genus (Fig. 1, Supplementary Fig. S4).

The coalescent trees were less sensitive to filtering than those inferred using concatenation (Fig. 2A), an observation previously made by other authors (Mitchell et al., 2017; Herrando-Moraira et al., 2018). This supports the ability of coalescence-based methods to effectively capture phylogenetic signals in the presence of noise and missing data, thus producing more robust results (Mirarab et al., 2016). With one exception, none of the filtering approaches increased branch support values significantly. Although filtering out fast-evolving sites greatly improved the support value of nodes associated with shorter branches under the concatenation approach, it is likely to lead to spurious results.

The notion that fast-evolving sites are responsible for incongruences in the results was supported by both gene concordance and site concordance factors (Supplementary Fig. S6A, B, E; Fig. S3). In general, gene concordance factors were low, even for nodes with a 100 % bootstrap value (Fig. 3A). The gCF tended to be lower than the sCF, suggesting that limited information from single locus trees and conflicting signals from biological processes all contributed to the observed discordance. The strong positive correlation between branch lengths and both concordance factors suggested that the heterogeneity between gene trees and species trees is mainly concentrated on the short branches (Supplementary Fig. S6) (Rosenberg, 2013).

A pattern of long terminal branches and shorter internal branches is typical for rapid radiations (Whitfield and Kjer, 2008; Bagley et al., 2020; Thomas et al., 2021). Short branches indicate short speciation intervals, which may not be enough for substitutions to accumulate, resulting in loci with low informativeness. Short intervals also increase the probability of ILS, which is a potential driver of incongruences. An influence of ILS was indicated both by the low site concordance factors and the distribution of theta values. For many nodes, we observed similar frequencies for the two alternative conformations of the quartet

(site discordance factor1 and 2, Fig. 3B) and high theta values, especially for nodes around the backbone of *Camellia* (Fig. 4B) where gene tree variation (Fig. 4D) and diversification rates (Yu et al., 2017; Cheng et al., 2022) are highest.

An alternative biological source of gene tree heterogeneity is the gene flow between species, such as through introgression and hybridization events. However, the level of gene flow in our dataset, as indicated by the percentage of asymmetrical triplets, was relatively low (2 %) in comparison to other challenging plant groups like the Malpighiales, where introgression among deep branches has been hypothesized (10 %) (Cai et al., 2021). We did observe a possibility of reticulated evolution within *Polyspora* in Southeast Asia, marking the first documented instance of this phenomenon. Further network analysis using Phylonet (Than et al., 2008) suggested the likelihood of a reticulation event between *Polyspora luzonica* and the common ancestor of *Polyspora marginata* and *Polyspora havilandii* (Note S2 and Figure S9). These species are distributed in the Malesia realm and were sampled in the nuclear phylogeny of the family for the first time. *Stewartia*, along with a few species in the *Camelliopsis* and *Camellia* sections of *Camellia,* also showed weak signals of introgression, as observed previously (Lin et al., 2019, Zhang et al., 2022, Zan et al., 2023). The cyto-nuclear discordance within these clades further supports this interpretation (Fig. 1).

Although these two biological processes contributed to the gene tree conflicts to some extent, the major contributor to low gene concordance is predicted to be gene tree estimation error for this dataset (Fig. 4). Applying variance decomposition analysis to simulated sequences with the same mutation rate, length, and topology, but without missing data, revealed that the frequency of gene tree estimation errors is predicted to count for 78 % of the 63 % explained gene tree variation. Although we have employed several well-recognized methods to mitigate gene tree estimation errors in our analysis, we observed that these methods did not substantially reduce the gene tree variation. On the one hand, gene tree estimation errors tend to be high for relatively short and less informative loci (Shen et al., 2020). Fig. 2B showed that the higher the average bootstrap of a gene tree, the closer it is to the average species tree, which suggests that part of the incongruence between gene trees was due to a lack of resolution (Xi et al., 2015). It is possible that this is linked to using a universal angiosperm probe set with variation in capture efficiency across different species, as the captured regions tend to be relatively conserved and alignments tend to be short after excluding regions with a lot of missing data. Similar situations have been reported in other phylogenetic studies using Angiosperm 353 probe set, such as in Dipsacales (Lee et al., 2021) and Artocarpeae (Moraceae) (Gardner, 2023). On the other hand, dataset with high levels of incomplete lineage sorting also tend to have more difficulties in "true" gene tree estimations than dataset with low levels of incomplete lineage sorting (Cai et al., 2021). The relatively high level of incomplete lineage sorting in the dataset further complicated the issue.

4.3. Conclusions and future directions

Using the universal probe set Angiosperm 353, we not only confirmed several intergeneric relationships previously inferred using plastid genomes, but also identified new relationships, even within the challenging *Camellia* genus. However, the overall support for the backbone of *Camellia* and tribe Theeae was low with high gene tree discordances. We identified clades associated with potential biological processes resulting from rapid diversification in the group, such as incomplete lineage sorting and introgression, apart from gene tree estimation errors that contribute to high heterogeneity within the gene tree dataset. Given that the reasons for gene tree discordances varied across different clades, further study could explore new phylogenetic method that balances between concatenation and coalescent approaches, thereby alleviating the impacts of both ILS and gene tree estimation errors (Smith et al., 2020). Simulation methods could also be employed to estimate the quantity of data needed to solve the most challenging relationships within the group.

We demonstrated that nuclear target enrichment is useful in providing valuable and novel insights into the evolutionary history of Theaceae. Its ability to extract information from historical and degraded samples complements the use of transcriptome/whole genome sequencing for modern samples. This is particularly important in clarifying the taxonomy of certain challenging groups, which requires molecular information from numerous individuals, spanning both historical specimens (ideally type specimens) and modern samples. However, the ability to fully resolve the recalcitrant nodes in Theaceae is relatively limited in the current dataset partly due to relatively short target length and low phylogenetic informativeness of genes, even after including intron regions. With the many transcriptomes and a few new complete genomes published for the group in recent years (Gong et al., 2022; Wu et al., 2022; Zan et al., 2023; Zhang et al., 2022), it worth expanding the universal probe set to include more variable, lineage-specific and function-related loci for future studies (McLay et al., 2021).

CRediT authorship contribution statement

Yujing Yan: Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization, Resources, Visualization, Writing – original draft, Writing – review & editing, Project administration. **Rute R. da Fonseca:** Methodology, Writing – review & editing. **Carsten Rahbek:** Supervision, Funding acquisition. **Michael K. Borregaard:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition. **Charles C. Davis:** Writing – review & editing, Supervision, Resources, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data generated in the study have been deposited in Dryad (https://doi.org/10.5061/dryad.5x69p8d85). Genbank accession numbers are provided in Supplementary Table S1.

Acknowledgments

We thank the Harvard University Herbaria and the New York Botanical Garden who generously provided the materials for the study. We thank the Bauer Core Facility of Harvard University for providing technical support during the laboratory process. We thank Lisa Pokorny, Lindsey Bechen and Elliot Gardner for offering valuable suggestions of the target enrichment lab work and Liming Cai for providing valuable advice and guidance for data analyses. The computations in this paper were run on the FASRC Odyssey cluster supported by the FAS Division of Science Research Computing Group at Harvard University. This work was supported by the Danish National Research Foundation through supporting the Center for Macroecology, Evolution and Climate [DNRF96, 2009-2019] to Yujing Yan, Rute R. da Fonseca, Carsten Rahbek, and Michael K. Borregaard; the Chinese Scholarship Council [No. 201606010394] and Harvard University Herbaria Postdoctoral Research Fellowship to Yujing Yan; setup funding from Harvard University to Charles C. Davis; and Carlsberg Young Researcher Award [CF19-0695] to Michael K. Borregaard.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2024.108089.

References

Andrews, S., 2010. FASTQC. A quality control tool for high throughput sequence data.

Avise, J.C., Robinson, T.J., 2008. Hemiplasy: A New Term in the Lexicon of Phylogenetics. Syst. Biol. 57, 503–507. https://doi.org/10.1080/10635150802164587.

Bagley, J.C., Uribe-Convers, S., Carlsen, M.M., Muchhala, N., 2020. Utility of targeted sequence capture for phylogenomics in rapid, recent angiosperm radiations: Neotropical Burmeistera bellflowers as a case study. Mol. Phylogenet. Evol. 152, 106769 https://doi.org/10.1016/j.ympev.2020.106769.

Baker, W.J., Dodsworth, S., Forest, F., Graham, S.W., Johnson, M.G., McDonnell, A., Pokorny, L., Tate, J.A., Wicke, S., Wickett, N.J., 2021. Exploring Angiosperms353: An open, community toolkit for collaborative phylogenomic research on flowering plants. Am. J. Bot. 108, 1059–1065. https://doi.org/10.1002/ajb2.1703.

Baker, W.J., Bailey, P., Barber, V., Barker, A., Bellot, S., Bishop, D., Botigué, L.R., Brewer, G., Carruthers, T., Clarkson, J.J., Cook, J., Cowan, R.S., Dodsworth, S., Epitawalage, N., Françoso, E., Gallego, B., Johnson, M.G., Kim, J.T., Leempoel, K., Maurin, O., Mcginnie, C., Pokorny, L., Roy, S., Stone, M., Toledo, E., Wickett, N.J., Zuntini, A.R., Eiserhardt, W.L., Kersey, P.J., Leitch, I.J., Forest, F., 2022. A comprehensive phylogenomic platform for exploring the angiosperm tree of life. Syst. Biol. 71, 301–319. https://doi.org/10.1093/sysbio/syab035.

Bieker, V.C., Martin, M.D., 2018. Implications and future prospects for evolutionary analyses of DNA in historical herbarium collections. Botany Letters 165, 409–418. https://doi.org/10.1080/23818107.2018.1458651.

Borowiec, M.L., 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. PeerJ 4, e1660. https://doi.org/10.7717/peerj.1660.

Brewer, G.E., Clarkson, J.J., Maurin, O., Zuntini, A.R., Barber, V., Bellot, S., Biggs, N., Cowan, R.S., Davies, N.M.J., Dodsworth, S., Edwards, S.L., Eiserhardt, W.L., Epitawalage, N., Frisby, S., Grall, A., Kersey, P.J., Pokorny, L., Leitch, I.J., Forest, F., Baker, W.J., 2019. Factors Affecting Targeted Sequencing of 353 Nuclear Genes From Herbarium Specimens Spanning the Diversity of Angiosperms. Frontiers. Plant Sci. 10.

Cai, L., Xi, Z., Lemmon, E.M., Lemmon, A.R., Mast, A., Buddenhagen, C.E., Liu, L., Davis, C.C., 2021. The perfect storm: gene tree estimation error, incomplete lineage sorting, and ancient gene flow explain the most recalcitrant ancient angiosperm clade, Malpighiales. Syst. Biol. 70, 491–507. https://doi.org/10.1093/sysbio/syaa083.

Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T., 2009. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348.

Chang, H.-T., 1998. Theaceae, in: Flora of Reipublicae Popularis Sinicae.

Chen, S., Zhou, Y., Chen, Y., Gu, J., 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

Cheng, L., Li, M., Han, Q., Qiao, Z., Hao, Y., Balbuena, T.S., Zhao, Y., 2022. Phylogenomics resolves the phylogeny of theaceae by using low-copy and multi-copy nuclear gene makers and uncovers a fast radiation event contributing to tea plants diversity. Biology 11, 1007. https://doi.org/10.3390/BIOLOGY11071007.

Cronn, R., Knaus, B.J., Liston, A., Maughan, P.J., Parks, M., Syring, J.V., Udall, J., 2012. Targeted enrichment strategies for next-generation plant biology. Am. J. Bot. 99, 291–311. https://doi.org/10.3732/ajb.1100356.

Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24, 332–340. https://doi.org/10.1016/j.tree.2009.01.009.

Gardner, E.M., 2023. Phylogenomic analyses of the Neotropical Artocarpeae (Moraceae) reveal a history of introgression and support the reinstatement of Acanthinophyllum. Mol. Phylogenet. Evol. 186, 107837 https://doi.org/10.1016/j.ympev.2023.107837.

Gong, W., Xiao, S., Wang, L., Liao, Z., Chang, Y., Mo, W., Hu, G., Li, W., Zhao, G., Zhu, H., Hu, X., Ji, K., Xiang, X., Song, Q., Yuan, D., Jin, S., Zhang, L., 2022. Chromosome-level genome of Camellia lanceoleosa provides a valuable resource for understanding genome evolution and self-incompatibility. Plant J. 110, 881–898. https://doi.org/10.1111/tpj.15739.

Groemping, U., Matthias, L., 2021. relaimpo: Relative importance of regression in linear models.

Grover, C.E., Salmon, A., Wendel, J.F., 2012. Targeted sequence capture as a powerful tool for evolutionary analysis [1]. Am. J. Bot. 99, 312–319. https://doi.org/10.3732/ajb.1100323.

Guo, J., Xu, W., Hu, Y., Huang, J., Zhao, Y., Zhang, L., Huang, C.-H., Ma, H., 2020. Phylotranscriptomics in Cucurbitaceae Reveal Multiple Whole-Genome Duplications and Key Morphological and Molecular Innovations. Mol. Plant 13, 1117–1133. https://doi.org/10.1016/j.molp.2020.05.011.

Herrando-Moraira, S., Calleja, J.A., Carnicero, P., Fujikawa, K., Galbany-Casals, M., Garcia-Jacas, N., Im, H.T., Kim, S.C., Liu, J.Q., López-Alvarado, J., López-Pujol, J., Mandel, J.R., Massó, S., Mehregan, I., Montes-Moreno, N., Pyak, E., Roquet, C., Sáez, L., Sennikov, A., Susanna, A., Vilatersana, R., 2018. Exploring data processing strategies in NGS target enrichment to disentangle radiations in the tribe Cardueae (Compositae). Mol. Phylogenet. Evol. 128, 69–87. https://doi.org/10.1016/j.ympev.2018.07.012.

Hillis, D.M., Heath, T.A., St. John, K., 2005. Analysis and visualization of tree space. Syst. Biol. 54, 471–482. https://doi.org/10.1080/10635150590946961.

Huang, H., Shi, C., Liu, Y., Mao, S.-Y., Gao, L.-Z., 2014. Thirteen Camellia chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. BMC Evol. Biol. 14, 151. https://doi.org/10.1186/1471-2148-14-151.

Jiang, B., Peng, Q.F., Shen, Z.G., Möller, M., Pi, E.X., Lu, H.F., 2010. Taxonomic treatments of Camellia (Theaceae) species with secretory structures based on integrated leaf characters. Plant Syst. Evol. 290, 1–20. https://doi.org/10.1007/s00606-010-0342-x.

Johnson, M.G., Gardner, E.M., Liu, Y., Medina, R., Goffinet, B., Shaw, A.J., Zerega, N.J. C., Wickett, N.J., 2016. HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. Appl. Plant Sci. 4 https://doi.org/10.3732/apps.1600016.

Johnson, M.G., Pokorny, L., Dodsworth, S., Botigué, L.R., Cowan, R.S., Devault, A., Eiserhardt, W.L., Epitawalage, N., Forest, F., Kim, J.T., Leebens-Mack, J.H., Leitch, I. J., Maurin, O., Soltis, D.E., Soltis, P.S., Wong, G.-K.-S., Baker, W.J., Wickett, N.J., 2019. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. Syst Biol 68, 594–606. https://doi.org/10.1093/sysbio/syy086.

Junier, T., Zdobnov, E.M., 2010. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics (oxford, England) 26, 1669–1670. https://doi.org/10.1093/bioinformatics/btq243.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780. https://doi.org/10.1093/molbev/mst010.

Kozlov, A.M., Darriba, D., Flouri, T., Morel, B., Stamatakis, A., 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. https://doi.org/10.1093/bioinformatics/btz305.

Kuhner, M.K., Yamato, J., 2015. Practical performance of tree comparison metrics. Syst. Biol. 64, 205–214. https://doi.org/10.1093/sysbio/syu085.

Lanfear, R., Frandsen, P.B., Wright, A.M., 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 msw260. 10.1093/molbev/msw260.

Larson, D.A., Walker, J.F., Vargas, O.M., Smith, S.A., 2020. A consensus phylogenomic approach highlights paleopolyploid and rapid radiation in the history of Ericales. Am. J. Bot. 107, 773–789. https://doi.org/10.1002/ajb2.1469.

Lee, A.K., Gilman, I.S., Srivastav, M., Lerner, A.D., Donoghue, M.J., Clement, W.L., 2021. Reconstructing Dipsacales phylogeny using Angiosperms353: issues and insights. Am. J. Bot. 108, 1122–1142. https://doi.org/10.1002/ajb2.1695.

Leebens-Mack, J.H., Barker, M.S., Carpenter, E.J., Deyholos, M.K., Gitzendanner, M.A., Graham, S.W., Grosse, I., Li, Z., Melkonian, M., Mirarab, S., Porsch, M., Quint, M., Rensing, S.A., Soltis, D.E., Soltis, P.S., Stevenson, D.W., Ullrich, K.K., Wickett, N.J., DeGironimo, L., Edger, P.P., Jordon-Thaden, I.E., Joya, S., Liu, T., Melkonian, B., Miles, N.W., Pokorny, L., Quigley, C., Thomas, P., Villarreal, J.C., Augustin, M.M., Barrett, M.D., Baucom, R.S., Beerling, D.J., Benstein, R.M., Biffin, E., Brockington, S. F., Burge, D.O., Burris, J.N., Burris, K.P., Burtet-Sarramegna, V., Caicedo, A.L., Cannon, S.B., Çebi, Z., Chang, Y., Chater, C., Cheeseman, J.M., Chen, T., Clarke, N. D., Clayton, H., Covshoff, S., Crandall-Stotler, B.J., Cross, H., DePamphilis, C.W., Der, J.P., Determann, R., Dickson, R.C., Di Stilio, V.S., Ellis, S., Fast, E., Feja, N., Field, K.J., Filatov, D.A., Finnegan, P.M., Floyd, S.K., Fogliani, B., García, N., Gâteblé, G., Godden, G.T., Goh, F. (Qi Y., Greiner, S., Harkess, A., Heaney, J.M., Helliwell, K.E., Heyduk, K., Hibberd, J.M., Hodel, R.G.J., Hollingsworth, P.M., Johnson, M.T.J., Jost, R., Joyce, B., Kapralov, M.V., Kazamia, E., Kellogg, E.A., Koch, M.A., Von Konrat, M., Könyves, K., Kutchan, T.M., Lam, V., Larsson, A., Leitch, A.R., Lentz, R., Li, F.W., Lowe, A.J., Ludwig, M., Manos, P.S., Mavrodiev, E., McCormick, M.K., McKain, M., McLellan, T., McNeal, J.R., Miller, R.E., Nelson, M.N., Peng, Y., Ralph, P., Real, D., Riggins, C.W., Ruhsam, M., Sage, R.F., Sakai, A.K., Scascitella, M., Schilling, E.E., Schlösser, E.M., Sederoff, H., Servick, S., Sessa, E.B., Shaw, A.J., Shaw, S.W., Sigel, E.M., Skema, C., Smith, A.G., Smithson, A., Stewart, C.N., Stinchcombe, J.R., Szövényi, P., Tate, J.A., Tiebel, H., Trapnell, D., Villegente, M., Wang, C.N., Weller, S.G., Wenzel, M., Weststrand, S., Westwood, J.H., Whigham, D. F., Wu, S., Wulff, A.S., Yang, Y., Zhu, D., Zhuang, C., Zuidof, J., Chase, M.W., Pires, J. C., Rothfels, C.J., Yu, J., Chen, C., Chen, L., Cheng, S., Li, J., Li, R., Li, X., Lu, H., Ou, Y., Sun, X., Tan, X., Tang, J., Tian, Z., Wang, F., Wang, J., Wei, X., Xu, X., Yan, Z., Yang, F., Zhong, X., Zhou, F., Zhu, Y., Zhang, Y., Ayyampalayam, S., Barkman, T.J., Nguyen, N. phuong, Matasci, N., Nelson, D.R., Sayyari, E., Wafula, E.K., Walls, R.L., Warnow, T., An, H., Arrigo, N., Baniaga, A.E., Galuska, S., Jorgensen, S.A., Kidder, T. I., Kong, H., Lu-Irving, P., Marx, H.E., Qi, X., Reardon, C.R., Sutherland, B.L., Tiley, G.P., Welles, S.R., Yu, R., Zhan, S., Gramzow, L., Theißen, G., Wong, G.K.S., 2019. One thousand plant transcriptomes and the phylogenomics of green plants. Nature 574, 679–685. 10.1038/s41586-019-1693-2.

Léveillé-Bourret, É., Starr, J.R., Ford, B.A., Moriarty Lemmon, E., Lemmon, A.R., 2018. Resolving rapid radiations within angiosperm families using anchored phylogenomics. Syst. Biol. 67, 94–112. https://doi.org/10.1093/sysbio/syx050.

Lin, H.-Y., Hao, Y.-J., Li, J.-H., Fu, C.-X., Soltis, P.S., Soltis, D.E., Zhao, Y.-P., 2019. Phylogenomic conflict resulting from ancient introgression following species diversification in Stewartia s.l. (Theaceae). Mol. Phylogenet. Evol. 135, 1–11. https://doi.org/10.1016/j.ympev.2019.02.018.

Mai, U., Mirarab, S., 2018. TreeShrink: Fast and accurate detection of outlier long branches in collections of phylogenetic trees. BMC Genomics 19. https://doi.org/10.1186/s12864-018-4620-2.

McLay, T.G.B., Birch, J.L., Gunn, B.F., Ning, W., Tate, J.A., Nauheimer, L., Joyce, E.M., Simpson, L., Schmidt-Lebuhn, A.N., Baker, W.J., Forest, F., Jackson, C.J., 2021. New targets acquired: Improving locus recovery from the Angiosperms353 probe set. Appl Plant Sci 9. https://doi.org/10.1002/aps3.11420.

Meleshko, O., Martin, M.D., Korneliussen, T.S., Schröck, C., Lamkowski, P., Schmutz, J., Healey, A., Piatkowski, B.T., Shaw, A.J., Weston, D.J., Flatberg, K.I., Szövényi, P., Hassel, K., Stenøien, H.K., 2021. Extensive genome-wide phylogenetic discordance is due to incomplete lineage sorting and not ongoing introgression in a rapidly radiated

bryophyte genus. Mol. Biol. Evol. 38, 2750–2766. https://doi.org/10.1093/molbev/msab063.
Meyer, B.S., Matschiner, M., Salzburger, W., 2017. Disentangling incomplete lineage sorting and introgression to refine species-tree estimates for lake tanganyika cichlid fishes. Syst. Biol. 66, 531–550. https://doi.org/10.1093/sysbio/syw069.
Min, T., Bartholomew, B., 2007. Theaceae. In: Flora of China, pp. 366–478.
Ming, T.L., 1999. A systematic synopsis of the genus *Camellia*. Acta Bot. Yunnanica 21, 149–159.
Mirarab, S., Bayzid, M.S., Warnow, T., 2016. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. Syst. Biol. 65, 366–380. https://doi.org/10.1093/sysbio/syu063.
Mitchell, N., Lewis, P.O., Lemmon, E.M., Lemmon, A.R., Holsinger, K.E., 2017. Anchored phylogenomics improves the resolution of evolutionary relationships in the rapid radiation of Protea L. Am. J. Bot. 104, 102–115. https://doi.org/10.3732/ajb.1600227.
Muñoz-Rodríguez, P., Carruthers, T., Wood, J.R.I., Williams, B.R.M., Weitemier, K., Kronmiller, B., Ellis, D., Anglin, N.L., Longway, L., Harris, S.A., Rausher, M.D., Kelly, S., Liston, A., Scotland, R.W., 2018. Reconciling conflicting phylogenies in the origin of sweet potato and dispersal to Polynesia. Curr. Biol. 28, 1246–1256.e12. https://doi.org/10.1016/j.cub.2018.03.020.
Murillo-A., J., Valencia-D., J., Orozco, C.I., Parra-O., C., Neubig, K.M., 2022. Incomplete lineage sorting and reticulate evolution mask species relationships in Brunelliaceae, an Andean family with rapid, recent diversification. American J of Botany 109, 1139–1156. 10.1002/ajb2.16025.
Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32, 268–274. https://doi.org/10.1093/molbev/msu300.
Paradis, E., Schliep, K., 2019. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35, 526–528. https://doi.org/10.1093/bioinformatics/bty633.
Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T.J., Manuel, M., Wörheide, G., Baurain, D., 2011. Resolving difficult phylogenetic questions: Why more sequences are not enough. PLoS Biol. 9 https://doi.org/10.1371/journal.pbio.1000602.
Rambaut, A., Grassly, N.C., 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13, 235–238. https://doi.org/10.1093/bioinformatics/13.3.235.
Rosenberg, N.A., 2013. Discordance of species trees with their most likely gene trees: a unifying principle. Mol. Biol. Evol. 30, 2709–2713. https://doi.org/10.1093/molbev/mst160.
Sand, A., Holt, M.K., Johansen, J., Brodal, G.S., Mailund, T., Pedersen, C.N.S., 2014. tqDist: a library for computing the quartet and triplet distances between binary or general trees. Bioinformatics 30, 2079–2080. https://doi.org/10.1093/bioinformatics/btu157.
Schmickl, R., Fishbein, M., Weitemier, K., McDonnell, A., Cronn, R.C., Straub, S.C.K., Liston, A., 2014. Hyb-Seq: combining target enrichment and genome skimming for plant phylogenomics. Appl. Plant Sci. 2, 1400042. https://doi.org/10.3732/apps.1400042.
Sealy, J.R., 1958. A Revision of the Genus Camellia., A Revision of the Genus Camellia. Royal Horticultural Society, London.
Shen, X.-X., Li, Y., Hittinger, C.T., Chen, X., Rokas, A., 2020. An investigation of irreproducibility in maximum likelihood phylogenetic inference. Nat. Commun. 11, 6096. https://doi.org/10.1038/s41467-020-20005-6.
Smith, S.A., Walker-Hale, N., Walker, J.F., Brown, J.W., 2020. Phylogenetic conflicts, combinability, and deep phylogenomics in plants. Syst. Biol. 69, 579–592. https://doi.org/10.1093/sysbio/syz078.
Smith, M.R., 2019. Quartet: comparison of phylogenetic trees using quartet and bipartition measures. 10.5281/zenodo.2536318.
Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics (oxford, England) 30, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033.
Straub, S.C.K., Parks, M., Weitemier, K., Fishbein, M., Cronn, R.C., Liston, A., 2012. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. Am. J. Bot. 99, 349–364. https://doi.org/10.3732/ajb.1100335.
Suh, A., Smeds, L., Ellegren, H., 2015. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. PLoS Biol. 13, e1002224. https://doi.org/10.1371/journal.pbio.1002224.
Than, C., Ruths, D., Nakhleh, L., 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinf. 9, 322. https://doi.org/10.1186/1471-2105-9-322.
Thomas, A.E., Igea, J., Meudt, H.M., Albach, D.C., Lee, W.G., Tanentzap, A.J., 2021. Using target sequence capture to improve the phylogenetic resolution of a rapid radiation in New Zealand Veronica. Am. J. Bot. 108, 1289–1306. https://doi.org/10.1002/ajb2.1678.
Vijayan, K., Zhang, W.J., Tsou, C.H., 2009. Molecular taxonomy of Camellia (Theaceae) inferred from nrITS sequences. Am. J. Bot. 96, 1348–1360. https://doi.org/10.3732/ajb.0800205.
Villaverde, T., Pokorny, L., Olsson, S., Rincón-Barrado, M., Johnson, M.G., Gardner, E. M., Wickett, N.J., Molero, J., Riina, R., Sanmartín, I., 2018. Bridging the micro- and macroevolutionary levels in phylogenomics: Hyb-Seq solves relationships from populations to species and above. New Phytol. 220, 636–650. https://doi.org/10.1111/nph.15312.
Wen, D., Yu, Y., Zhu, J., Nakhleh, L., 2018. Inferring Phylogenetic Networks Using PhyloNet. Syst. Biol. 67, 735–740. https://doi.org/10.1093/sysbio/syy015.
Whitfield, J.B., Kjer, K.M., 2008. Ancient rapid radiations of insects: challenges for phylogenetic analysis. Annu. Rev. Entomol. 53, 449–472. https://doi.org/10.1146/annurev.ento.53.103106.093304.
Wu, Q., Tong, W., Zhao, H., Ge, R., Li, R., Huang, J., Li, F., Wang, Y., Mallano, A.I., Deng, W., Wang, W., Wan, X., Zhang, Z., Xia, E., 2022. Comparative transcriptomic analysis unveils the deep phylogeny and secondary metabolite evolution of 116 Camellia plants. Plant J. 111, 406–421. https://doi.org/10.1111/tpj.15799.
Xi, Z., Liu, L., Davis, C.C., 2015. Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased. Mol. Phylogenet. Evol. 92.
Xia, E.H., Zhang, H.B., Sheng, J., Li, K., Zhang, Q.J., Kim, C., Zhang, Y., Liu, Y., Zhu, T., Li, W., Huang, H., Tong, Y., Nan, H., Shi, C., Shi, C., Jiang, J.J., Mao, S.Y., Jiao, J.Y., Zhang, D., Zhao, Y., Zhao, Y.J., Zhang, L.P., Liu, Y.L., Liu, B.Y., Yu, Y., Shao, S.F., Ni, D.J., Eichler, E.E., Gao, L.Z., 2017. The Tea Tree Genome Provides Insights into Tea Flavor and Independent Evolution of Caffeine Biosynthesis. Mol. Plant 10, 866–877. https://doi.org/10.1016/j.molp.2017.04.002.
Yan, Y., Davis, C.C., Dimitrov, D., Wang, Z., Rahbek, C., Borregaard, M.K., 2021. Phytogeographic history of the Tea family inferred through high-resolution phylogeny and fossils. Syst. Biol. 2–47 https://doi.org/10.1093/sysbio/syab042.
Yang, J.B., Yang, S.X., Li, H.T., Yang, J., Li, D.Z., 2013. Comparative chloroplast genomes of camellia species. PLoS One 8. https://doi.org/10.1371/journal.pone.0073053.
Yang, Z., 2006. Phylogeny reconstruction: overview, in: Yang, Z. (Ed.), Computational Molecular Evolution. Oxford University Press, p. 0. 10.1093/acprof:oso/9780198567028.003.0003.
Yu, X.Q., Gao, L.M., Soltis, D.E., Soltis, P.S., Yang, J.B., Fang, L., Yang, S.X., Li, D.Z., 2017. 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. https://doi.org/10.1111/nph.14683.
Zan, T., He, Y.-T., Zhang, M., Yonezawa, T., Ma, H., Zhao, Q.-M., Kuo, W.-Y., Zhang, W.-J., Huang, C.-H., 2023. Phylogenomic analyses of Camellia support reticulate evolution among major clades. Mol. Phylogenet. Evol. 182, 107744 https://doi.org/10.1016/j.ympev.2023.107744.
Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S., 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinf. 19, 153. https://doi.org/10.1186/s12859-018-2129-y.
Zhang, C., Huang, C., Liu, M., Hu, Y., Panero, J.L., Luebert, F., Gao, T., Ma, H., 2021. Phylotranscriptomic insights into Asteraceae diversity, polyploidy, and morphological innovation. JIPB 63, 1273–1293. https://doi.org/10.1111/jipb.13078.
Zhang, Q., Zhao, L., Folk, R.A., Zhao, J.-L., Zamora, N.A., Yang, S.-X., Soltis, D.E., Soltis, P.S., Gao, L.-M., Peng, H., Yu, X.-Q., 2022. Phylotranscriptomics of Theaceae: generic level relationships, reticulation and whole-genome duplication. Ann. Bot. 1–14 https://doi.org/10.1093/aob/mcac007.
Zhao, D.-W., Hodkinson, T.R., Parnell, J.A.N., 2023. Phylogenetics of global Camellia (Theaceae) based on three nuclear regions and its implications for systematics and evolutionary history. J. System. Evol. https://doi.org/10.1111/jse.12837.