Phylogenomic analysis of the hemp family (Cannabaceae) reveals deep cyto-Nuclear discordance and provides new insights into generic relationships

Journal of Systematics and Evolution
Fu, X.; Liu, G.; Velzen, R.; Stull, G.W.; Tian, Q. et al
https://doi.org/10.1111/jse.12920

JSE Journal of Systematics and Evolution

doi: 10.1111/jse.12920

Research Article

Phylogenomic analysis of the hemp family (Cannabaceae) reveals deep cyto-nuclear discordance and provides new insights into generic relationships

Xiao-Gang Fu[1,2,3†], Shui-Yin Liu[2,3†], Robin van Velzen[4], Gregory W. Stull[2], Qin Tian[2,3], Yun-Xia Li[2,3], Ryan A. Folk[5], Robert P. Guralnick[6], Heather R. Kates[6], Jian-Jun Jin[7], Zhong-Hu Li[1], Douglas E. Soltis[6,8], Pamela S. Soltis[6], and Ting-Shuang Yi[2,3*]

[1]Key Laboratory of Resource Biology and Biotechnology in Western China, Ministry of Education, College of Life Sciences, Northwest University, Xi'an 710069, China
[2]Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, China
[3]University of Chinese Academy of Sciences, Beijing 100049, China
[4]Biosystematics Group, Wageningen University and Research Center, Wageningen 6708 PB, the Netherlands
[5]Department of Biological Sciences, Mississippi State University, Starkville, MS 39762, USA
[6]Florida Museum of Natural History, University of Florida, Gainesville, FL 32611, USA
[7]Department of Ecology, Evolution, and Environmental Biology, Columbia University, New York, NY 10027, USA
[8]Department of Biology, University of Florida, Gainesville, FL 32611, USA
[†]These authors contributed equally to this work.
*Author for correspondence. E-mail: tingshuangyi@mail.kib.ac.cn
Received 6 June 2022; Accepted 27 September 2022; Article first published online 30 September 2022

Abstract Cannabaceae are a relatively small family of angiosperms, but they include several species of huge economic and cultural significance: marijuana or hemp (*Cannabis sativa*) and hops (*Humulus lupulus*). Previous phylogenetic studies have clarified the most deep relationships in Cannabaceae, but relationships remain ambiguous among several major lineages. Here, we sampled 82 species representing all genera of Cannabaceae and utilized a new dataset of 90 nuclear genes and 82 chloroplast loci from Hyb-Seq to investigate the phylogenomics of Cannabaceae. Nuclear phylogenetic analyses revealed a robust and consistent backbone for Cannabaceae. We observed nuclear gene-tree conflict at several deep nodes in inferred species trees, also cyto-nuclear discordance concerning the relationship between *Gironniera* and *Lozanella* and the relationships among *Trema* s.l. (including *Parasponia*), *Cannabis* + *Humulus*, and *Chaetachme* + *Pteroceltis*. Coalescent simulations and network analyses suggest that observed deep cyto-nuclear discordances were most likely to stem from incomplete lineage sorting (ILS); nuclear gene-tree conflict might be caused by both ILS and gene flow between species. All genera of Cannabaceae were recovered as monophyletic, except for *Celtis*, which consisted of two distinct clades: *Celtis* I (including most *Celtis* species) and *Celtis* II (including *Celtis gomphophylla* and *Celtis schippii*). We suggest that *Celtis* II should be recognized as the independent genus *Sparrea* based on both molecular and morphological evidence. Our work provides the most comprehensive and reliable phylogeny to date for Cannabaceae, enabling further exploration of evolutionary patterns across this family and highlighting the necessity of comparing nuclear with chloroplast data to examine the evolutionary history of plant groups.

Key words: ancient hybridization, Cannabaceae, *Celtis*, classification, cyto-nuclear discordance, incomplete lineage sorting, phylogenomics, *Sparrea*.

1 Introduction

Cannabaceae comprise ca. 117 species distributed in tropical, subtropical, and temperate regions of the world (Jin et al., 2020). Members of Cannabaceae are mostly trees or shrubs, rarely vines (*Humulus*) or erect herbs (*Cannabis*), and show considerable diversity in morphology and habitat (Yang et al., 2013). The two most economically important plants in this family are marijuana or hemp (*Cannabis sativa* L.) and hops (*Humulus lupulus* L.). Hemp was one of the first plants to be domesticated, probably during early Neolithic times in China, and has been used for thousands of years as a source of fiber, food, and medicine (Zlas et al., 1993; Pringle, 1997; Kovalchuk et al., 2020; Ren et al., 2021). The

female inflorescences of hops have been an essential ingredient in beer brewing since the early Middle Ages (Behre, 1999; Biendl & Pinzl, 2007; Zanoli & Zavatti, 2008). Another economic species is wingceltis (*Pteroceltis tatarinowii* Maxim.), whose bark fiber represents the key material for making traditional Chinese Xuan paper (Li et al., 2012).

The phylogenetic placement and circumscription of Cannabaceae have changed considerably following molecular studies. This family is a member of the "urticalean rosid" clade, a morphologically well defined group now included in a more broadly circumscribed Rosales (Sytsma et al., 2002; Soltis et al., 2011; Zhang et al., 2011; Sun et al., 2016). The circumscription and classification of Cannabaceae have long been controversial (see Table S1). Originally including only *Cannabis* and *Humulus* (Rendle, 1925), this family now comprises eight additional genera (*Aphananthe*, *Celtis*, *Chaetachme*, *Gironniera*, *Lozanella*, *Parasponia*, *Pteroceltis*, and *Trema*) formerly classified as Celtidaceae or Celtidoideae within Ulmaceae (Link, 1831; Engler & Prantl, 1893). This expanded concept of Cannabaceae, first proposed by Wiegrefe et al. (1998), is supported by multiple molecular studies (i.e., Song et al., 2001; Sytsma et al., 2002; Sattarian, 2006; Yang et al., 2013; Sun et al., 2016; van Velzen et al., 2018; Zhang et al., 2018b; Jin et al., 2020). Several studies have also shown that *Parasponia* is nested within *Trema* sensu stricto (s.s.) with weak support (i.e., Yesson et al., 2004; Sattarian, 2006; Yang et al., 2013; Sun et al., 2016; Jin et al., 2020) or with limited sampling (i.e., Zhang et al., 2018b). Accordingly, species of the former genus were recently subsumed into the latter (Christenhusz et al., 2018).

Relationships within Cannabaceae have been greatly clarified by phylogenetic studies over the last 30 years using chloroplast, mitochondrial, and/or nuclear ribosomal DNA sequences (Fig. S1). *Aphananthe* has been consistently recovered as sister to the rest of the family, with *Gironniera* and *Lozanella* either being successive sisters or forming a clade that is sister to the remaining genera, which together form a quadripartite clade including *Celtis*, *Cannabis* + *Humulus*, *Chaetachme* + *Pteroceltis*, and *Trema* sensu lato (s.l.) (Song et al., 2001; Song & Li, 2002; Sytsma et al., 2002; Sattarian, 2006; Yang et al., 2013; Sun et al., 2016; van Velzen et al., 2018; Zhang et al., 2018b; Jin et al., 2020). However, previous phylogenetic studies of the family have included limited sampling of either DNA markers (i.e., Sattarian, 2006; Yang et al., 2013; Sun et al., 2016; Jin et al., 2020) or species (i.e., Zhang et al., 2018b), with a relatively limited perspective overall from the nuclear genome. Despite continued progress, major outstanding questions remain, including the positions of *Gironniera* and *Lozanella*, major relationships within the quadripartite clade, and the overall congruence of phylogenetic relationships inferred from plastid vs. nuclear datasets.

To address these questions, we generated a new phylogenomic dataset for the family including 90 nuclear genes (for 82 ingroup species) and 82 chloroplast loci (for 90 ingroup species), to resolve both deep and shallow relationships in Cannabaceae. We obtained a well resolved nuclear phylogenetic framework and clarified several contentious relationships in Cannabaceae. We also explored patterns of gene-tree and cyto-nuclear (i.e., chloroplast–nuclear) discordance across Cannabaceae phylogeny to identify the potential biological sources—for example, ancient hybridization vs. incomplete lineage sorting (ILS)—of the observed phylogenomic conflict. Based on our robust phylogenetic framework and a survey of relevant morphological traits, we propose a revised generic classification of the family, including the renewed recognition of *Sparrea*.

2 Material and Methods

2.1 Taxon sampling and target enrichment sequencing

We sampled 83 species of Cannabaceae (ca. 70% of accepted species according to Jin et al., 2020), representing all genera of Cannabaceae with 67%–100% species sampling for each genus. Two species from families closely related to Cannabaceae (Zhang et al., 2011; Jin et al., 2020) were selected as the outgroups: *Broussonetia papyrifera* (L.) L'Hér. ex Vent. from Moraceae and *Debregeasia saeneb* (Forssk.) Hepper & J.R.I.Wood from Urticaceae. Leaf material for the 86 samples was obtained from the San Francisco Botanical Garden and the following herbaria: A, CAS, K, KUN, L, MO, NY, OS, RSA, and TEX (acronyms following Index Herbariorum). Detailed sampling information is listed in Table S2.

Total genomic DNA was extracted from herbarium or silica-dried material with a modified cetyl trimethyl ammonium bromide (CTAB) protocol (Doyle & Doyle, 1987; Folk et al., 2021). For target enrichment sequencing (hereafter Hyb-Seq), we used an exonic bait set for 100 single-copy or low-copy genes (133 433 bp in total) developed for phylogenetic analyses across the rosid clade (Folk et al., 2021), which includes Cannabaceae. This bait set was developed using 78 rosid transcriptomes and *Arabidopsis thaliana* (L.) Heynh. as a genome reference. In addition to the 100 nuclear loci, we obtained 82 off-target chloroplast loci from the Hyb-Seq data (see more details below). Genomic DNA quantification, library preparation, target enrichment, and Illumina sequencing with 150-bp paired-end reads were conducted by Rapid Genomics (Gainesville, Florida, USA).

2.2 Read filtering, assembly, and processing

2.2.1 Nuclear dataset

The raw paired-end reads obtained from Hyb-Seq were filtered using Trimmomatic version 0.36 (Bolger et al., 2014) to remove adapters and low-quality bases (Phred score = 33; ILLUMINACLIP: TruSeq3-PE-adapters.fa:2:30:10:8:TRUE; SLIDINGWINDOW:20:20). We used HybPiper version 1.3.1 (Johnson et al., 2016) to assemble the cleaned reads and for post-processing (https://github.com/mossmatters/HybPiper/), using as a reference 100 protein sequences from *A. thaliana*, which correspond to the gene sequences for probe design. More specifically, we used the wrapper script "reads_first.py" to map reads to each reference protein sequence using BLASTX version 2.7.1 (Camacho et al., 2009). The binned reads were assembled separately for each gene using SPAdes version 3.12.0 (Bankevich et al., 2012). Then the assembled contigs were aligned to the reference protein sequences using Exonerate version 2.4.0 (Slater & Birney, 2005). We generated a heatmap to show the recovery efficiency for each targeted gene and sample using the scripts "get_seq_lengths.py" and

"gene_recovery_heatmap_ggplot.R." An average of 81 (4–92) targeted nuclear sequences was recovered for each sample. One sample with poor assembly results (*Celtis balansae* Planch., with <10 recovered sequences, sequence lengths <25% the length of the complete targeted gene, and aberrant sequences compared with those of its relatives) was excluded from subsequent analyses, while another sample (i.e., *Celtis rubrovenia* Elmer) with four assembled genes was retained to include as many species as possible in this study.

For orthology inference, we first used the HybPiper scripts "paralog_investigator.py" and "paralog_retriever.py" to retrieve all assembled exons (>85% of the target length) for each gene of each species. Based on this approach, only a small number of genes showed possible evidence of paralogues (i.e., two or more sequences per gene for any particular species), but automated scripts were not always successful at identifying the correct orthologue upon inspection. Accordingly, two strategies were applied to paralogue processing. First, we directly excluded from each gene any species/sample with two or more copies, which might be the most certain way to exclude paralogues, without resulting in too much missing data (due to two species per gene and two genes per species on average with possible paralogues; see Supplementary Materials for this statistical result). Second, we aligned each gene sequence with potential paralogues using MAFFT version 7.305b (Katoh & Standley, 2013) under default parameters and inspected results in GENEIOUS version 8.0.2 (Kearse et al., 2012). Specifically, we visualized this gene alignment (containing possible paralogues), compared sequences of samples with those of their respective closely related species, and for each sample finally retained as the putative orthologue the sequence having the highest percentage identity with the sequences of its relatives.

After paralogue processing, each gene alignment was further pruned to remove poorly aligned columns using trimAl version 1.2 (Capella-Gutiérrez et al., 2009) with the "-automated1" option. Additionally, alignments missing >75% of the sampled species were excluded from subsequent analyses. This resulted in two final nuclear datasets (corresponding to the two strategies of paralogue processing described above), both of which included 90 genes and 84 species. Given congruent phylogenetic results of both concatenation and coalescent-based methods (except for several infrageneric relationships) based on these two nuclear datasets, only the dataset from the second strategy of paralogue processing were used for subsequent analyses (see Table S3 for further information on each gene).

2.2.2 Chloroplast dataset

Chloroplast DNA sequences were assembled from off-target reads following the same methods described above for the nuclear dataset except that BWA version 0.7.15 (Li & Durbin, 2009) was used for read mapping. We assembled 78 chloroplast protein-coding gene sequences and four non-coding sequences (*atpB-rbcL*, *psbA-trnH*, *trnL-trnF*, and *rps16* intron), using extracted loci from *Celtis biondii* Pamp. as the reference. We also extracted these regions from another 23 annotated plastomes of Cannabaceae species available in GenBank (see Table S4 for the accession numbers) using the python script "get_annotated_regions_from_gb.py" in Zhang et al. (2020) (https://github.com/Kinggerm/PersonalUtilities). Sequences from these different data sources—that is, assembled from Hyb-Seq data, extracted from annotated plastomes, and directly downloaded from GenBank (all accession numbers listed in Table S5)—were then combined by locus, and each locus was independently aligned and filtered using the same method as applied to the nuclear genes. This resulted in a dataset of 82 chloroplast loci and 92 species for subsequent phylogenetic analyses (Table S6).

2.3 Nuclear phylogenetic analysis

We reconstructed nuclear phylogenies of Cannabaceae using concatenation and coalescent-based approaches. We first concatenated the cleaned nuclear gene alignments into a supermatrix using the script "concatenate_fasta.py" in Zhang et al. (2020) (https://github.com/Kinggerm/PersonalUtilities) for ML analyses. The concatenated ML tree was inferred with the edge-linked partition model (i.e., branch lengths linked between different partitions; see Supplementary Materials for the partitions and substitution models) using IQ-TREE version 1.6.12 (Nguyen et al., 2015); branch support was estimated using the SH-like approximate likelihood ratio test (SH-aLRT) (Guindon et al., 2010) as well as ultrafast bootstrapping (UFBoot) with 1000 replicates (Hoang et al., 2018). We also inferred a concatenated ML tree using RAxML version 8.2.12 (Stamatakis, 2014) under (A) an unpartitioned GTR-GAMMA model and (B) a GTR-GAMMA model partitioned according to the best-fitting partitioning scheme (with 43 partitions provided in Supplementary Materials) inferred by PartitionFinder version 2.1.1 (Lanfear et al., 2017). The latter ML inference from RAxML was used for subsequent molecular dating and network analyses. To examine the effect of missing data on our phylogenetic results, we also conducted RAxML analyses under the GTR-GAMMA model with 27 partitions determined by PartionFinder, based on a concatenated supermatrix of 49 nuclear gene alignments with >90% of the sampled species (see Supplementary Materials for this optimal partitioning scheme on this reduced supermatrix). Branch support in all concatenated RAxML analyses was estimated with 1000 bootstrap replicates.

Coalescent-based species-tree analyses were performed in ASTRAL-III version 5.6.3 (Zhang et al., 2018a) including gene trees of the 90 nuclear loci inferred by RAxML with the GTR-GAMMA model and 200 rapid bootstraps. Branch support values were estimated with local posterior probabilities (LPPs; Sayyari & Mirarab, 2016). To accommodate poorly supported branches in the gene trees, ASTRAL analyses were performed in two ways: (A) by directly using gene trees inferred by RAxML and (B) by using gene trees with poorly supported branches (i.e., BS support <10%) collapsed prior to the analysis using the "nw_ed" program in Newick Utilities (Junier & Zdobnov, 2010), which was shown to improve the accuracy of tree inference (Zhang et al., 2017). Results from the latter ASTRAL analysis were used in subsequent conflict analyses, coalescent simulations, and network analyses. All nuclear species trees were visualized using FigTree version 1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/) and Adobe Illustrator (AI) 2020 software.

2.4 Chloroplast phylogenetic analysis

We inferred the chloroplast ML tree for Cannabaceae based on the concatenated alignment of 82 chloroplast loci using

RAxML under a partitioned GTR-GAMMA model. The optimal partitioning scheme across the concatenated chloroplast alignment was inferred using the corrected Akaike Information Criterion (AICc) in PartitionFinder (with 32 partitions provided in Supplementary Materials). Branch support values were estimated through 1000 fast bootstrap replicates with the option "-f a." We also inferred the chloroplast ML tree with a concatenated alignment using IQ-TREE following the same approach used for the nuclear dataset (see Supplementary Materials for the partitions and substitution models). ASTRAL was not applied to the chloroplast dataset, given that the entire chloroplast genome is usually considered a single locus sharing the same genealogy (Doyle, 1992; Gitzendanner et al., 2018; Doyle, 2022; but see Walker et al., 2019). The chloroplast ML trees were visualized using FigTree and AI.

2.5 Divergence time estimation

To establish a timeframe for ILS and/or gene flow events responsible for the observed deep cyto-nuclear and gene-tree discordances, we conducted dating analyses using penalized likelihood implemented in treePL (Smith & O'Meara, 2012), with the best RAxML tree inferred by the nuclear dataset as input. We used the same five well vetted fossils used by Jin et al. (2020) (Table S7) for calibration. The optimal smoothing value of the final treePL analysis was determined by the lowest χ^2 value through cross-validation tests. Following the empirical guide by Maurin (2020), we performed treePL analyses on 1000 bootstrap replicates (with the topology fixed to the best RAxML tree obtained in our nuclear gene analyses, but with branch lengths allowed to vary) to estimate confidence intervals on the inferred ages. We used TreeAnnotator version 2.6.3 (Bouckaert et al., 2014) to map confidence intervals for all node ages on the dated nuclear ML tree.

We additionally conducted a Bayesian dating analysis using BEAST version 2.6.4 (Bouckaert et al., 2014) under an uncorrelated lognormal relaxed clock. Given the extensive computational demands of BEAST analyses and the possible impact of nuclear gene conflict on divergence time estimation (Mendes & Hahn, 2016), we identified the 20 most clock-like nuclear genes using SortaDate (Smith et al., 2018) based on three criteria (less topological conflict with the best nuclear RAxML tree, lower root-to-tip variance [i.e., more clock-likeness], and more discernible informative sites). The BEAST analysis was performed under a birth–death process and GTR-GAMMA substitution model. We fixed the topology to the best nuclear RAxML tree and employed the same five fossil calibrations used in the treePL analysis. The Monte Carlo Markov chain was run for 2000 million generations sampling every 1000 generations. The first 25% of posterior trees were discarded as burn-in, and each parameter was checked to assure an effective sample size greater than 200 for convergence using Tracer version 1.7.1 (Rambaut et al., 2018). Finally, a maximum clade credibility tree with median heights was generated using post-burn-in posterior trees in TreeAnnotator.

2.6 Conflict analyses

We investigated topological concordance and conflict among the nuclear gene trees using phyparts version 0.0.1 (Smith et al., 2015). Given a main input topology, this method summarizes for each branch the numbers of genes supporting this bipartition, the number of genes supporting a main alternative bipartition, the number of genes supporting other alternative bipartitions, and the number of genes with no information (due to low support or insufficient sampling). We conducted two analyses, mapping the 90 nuclear genes against (A) the nuclear ASTRAL species tree and (B) the chloroplast ML tree. The latter was done to determine whether clades recovered in the chloroplast topology are supported by subsets of the nuclear genome in cases of deep conflict between the main chloroplast and nuclear phylogenies. Before conducting the conflict analyses, all trees were rooted with *Broussonetia papyrifera* and *Debregeasia saeneb* as outgroups using the "pxrr" program in Phyx (Brown et al., 2017). For gene trees with the outgroups missing, we used *Aphananthe* species for rooting (as this genus is strongly supported as sister to the rest of the family). To focus on well supported gene-tree conflict in Cannabaceae, we considered all branches with BS < 70% to be uninformative, following previous studies (i.e., Stubbs et al., 2020; Stull et al., 2020). The phyparts results were visualized with the script "phypartspiecharts.py" (https://github.com/mossmatters/phyloscripts/tree/master/phypartspiecharts).

2.7 Coalescent simulations and network analysis

We used a coalescent simulation strategy to explore whether ILS alone could generate observed instances of cyto-nuclear discordance in Cannabaceae, following several studies (i.e., Folk et al., 2017; Stull et al., 2020; Wang et al., 2021). Specifically, we simulated 1000 chloroplast trees under the coalescent model using the modified script "generateCoalescentTrees.py" (https://github.com/yongzhiyang2012/Chloranthus-sessilifolius-genome) based on the library DendroPy version 4.1.0 (Sukumaran & Holder, 2010). The nuclear ASTRAL species tree was used as the guide tree, with branch lengths rescaled by a factor of two and four to account for organellar inheritance (McCauley, 1994). We then mapped all simulated chloroplast trees to the inferred chloroplast tree using phyparts. The number of simulated trees supporting or in conflict with each branch of the chloroplast topology was visualized using the script "phypartspiecharts.py." At branches with cyto-nuclear discordance, large numbers of simulated trees supporting the chloroplast topology indicate that ILS alone is likely sufficient to explain the observed conflict. In contrast, if relatively few or no simulated organellar trees support the chloroplast topology, this suggests that the observed chloroplast topology is not expected given the distribution of simulated chloroplast trees and that another explanation (in particular, gene flow or chloroplast capture) is necessary to explain the observed cyto-nuclear discordance (García et al., 2017; Wang et al., 2021).

We also inferred phylogenetic networks using PhyloNet version 3.8.0 (Than et al., 2008; Yu et al., 2012; Wen et al., 2018) to explore the possibility of reticulation in the evolutionary history of Cannabaceae. Network approaches are advantageous in that they can account for both gene flow and ILS (Cao et al., 2019), but their computational complexity limits the number of species or individuals that can be included in a single analysis. Because our main focus

here is on possible deep reticulation in Cannabaceae (due to the observed cyto-nuclear discordances involving the phylogenetic placement of genera of Cannabaceae), we reduced our taxon sampling for these analyses to 11 ingroup species representing all nine genera of Cannabaceae (one representative per genus except for *Celtis*, for which we used a representative for each of the three clades recovered by our study) and two outgroup species. The representatives selected for each genus (or clade) were those with the greatest gene occupancy. The resulting dataset for network analyses (hereafter referred to as the "11-taxon dataset") included 85 nuclear genes (as we excluded five genes with fewer than seven representative species). We conducted network searches allowing one to five reticulation events using the "InferNetwork_MPL" command, with a 50% BS threshold for the gene-tree branches following several studies (i.e., Wang et al., 2021; Morales-Briones & Kadereit, 2022; Sun et al., 2022). To test whether a bifurcating tree can better fit the gene trees, the log likelihood scores of strictly bifurcating trees (the nuclear RAxML tree, the nuclear ASTRAL tree, and the chloroplast ML tree, all pruned to match these representative species) were also computed with the same set of gene trees using the "CalGTProb" command. The optimal network was finally determined by the AIC, the AICc, and the Bayesian Information Criterion (BIC), following the calculation method of Morales-Briones et al. (2021). To assess the impact of taxon sampling, we additionally performed the same PhyloNet analysis using two additional datasets that included 82 and 83 pruned nuclear gene trees, respectively. These gene trees were generated by randomly sampling one representative per genus except for *Celtis*, for which we included three representatives, as noted above. These additional datasets are hereafter referred to as the "Random1-11-taxon dataset" and "Random2-11-taxon dataset." All phylogenetic networks were visualized using the Julia package PhyloPlots (https://github.com/cecileane/PhyloPlots.jl).

2.8 Data accessibility

The resulting nuclear and chloroplast gene alignments, phylogenetic trees, and major results from all analyses can be found in Dryad: https://doi.org/10.5061/dryad.hmgqnk9mc.

3 Results

3.1 Data assembly

The recovery efficiencies for the 100 targeted nuclear genes and the 82 chloroplast loci from the Hyb-Seq data are shown using heatmaps (Figs. S2A, S2B). The number of assembled nuclear genes for each sample ranged from four in *C. rubrovenia* to 92 in *Trema guineensis* (Schumach. & Thonn.) Ficalho and *Trema tomentosa* (Roxb.) H. Hara, with an average of 81 nuclear genes recovered (Table S8). We found that only 36 of the 100 nuclear genes showed paralogy warnings. For each of these genes, possible paralogues were only observed in ca. two samples on average (see Supplementary Materials for this statistical result). Our final concatenated nuclear matrix (after cleaning) had an aligned length of 99 954 bp. From the off-target reads, we assembled chloroplast loci for 80 species. More information on the assembled chloroplast loci is available in Table S9. The final concatenated chloroplast matrix (after cleaning) had an aligned length of 74 912 bp (Table 1).

3.2 Phylogenetic analyses

The nuclear phylogenies inferred from concatenation and coalescent-based approaches were strongly supported and overwhelmingly congruent with each other, except for some infrageneric relationships (Figs. 1, S3, S4). Notably, the different levels of missing data, partitioning strategies, and tree inference methods (Figs. 1, S5, S6, S7) had little effect on inferred relationships in Cannabaceae. The monophyly of each genus in Cannabaceae was fully supported (BS = 100%; UFBoot = 100%; LPP = 1.00) except for *Celtis*, which was resolved into two separate clades across all nuclear analyses: *Celtis* I (including most *Celtis* species) immediately sister to all other genera of the quadripartite clade, and *Celtis* II (including *Celtis gomphophylla* and *Celtis schippii*) sister to *Celtis* I + the remainder of the quadripartite clade. More broadly in Cannabaceae, *Aphananthe* was recovered as sister to the rest of the family without exception. *Gironniera*, *Lozanella*, and *Celtis* II were successive sisters to the quadripartite clade comprising the remaining Cannabaceae genera (*Celtis* I, *Cannabis* + *Humulus*, *Chaetachme* + *Pteroceltis*, and *Trema* s.l.). Within the quadripartite clade, *Cannabis* + *Humulus* was supported as sister to *Chaetachme* + *Pteroceltis* (hereafter the CHCP subclade) (BS = 92%; UFBoot = 97%; LPP = 0.57). The CHCP subclade was sister to *Trema* s.l., and they together formed a clade sister to *Celtis* I (BS = 100%; UFBoot = 100%; LPP = 1.00). The *Parasponia* clade was nested in *Trema* s.l. with weak to moderate support (BS = 53%; UFBoot = 89%; LPP = 0.87).

The topologies of the chloroplast ML trees inferred using RAxML and IQ-TREE (Fig. 2) were generally consistent with those of the nuclear trees (including the two separate clades of *Celtis*), but two deep nodes with cyto-nuclear discordance were observed (Fig. 3). In our chloroplast ML tree, *Gironniera* was recovered as sister to *Lozanella* (BS = 53%; UFBoot = 92%). Within the quadripartite clade, *Trema* s.l. was recovered as sister to *Cannabis* + *Humulus* (BS = 100%; UFBoot = 100%), and they together formed a clade sister to *Chaetachme* + *Pteroceltis* (BS = 82%; UFBoot = 94%).

Table 1 A summary of the nuclear and chloroplast datasets used in this study

Dataset	Number of species	Number of loci	Length of concatenated alignment (bp)	Percentage of gaps in concatenated alignment	Substitution model used in ML analysis
Nuclear	84	90	99 954	20.86	GTR-GAMMA
Chloroplast	90	82	74 912	51.30	GTR-GAMMA

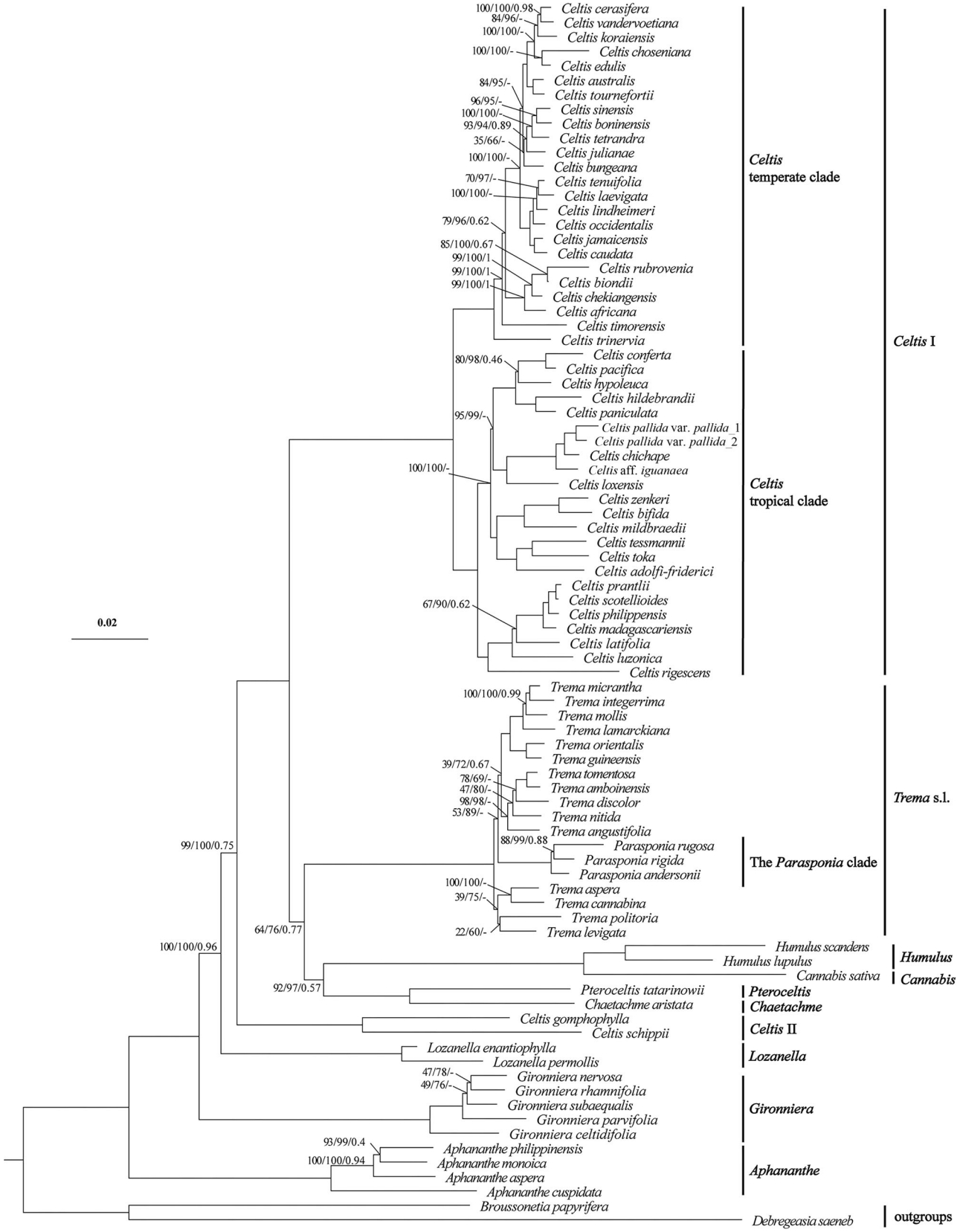

Fig. 1. Continued

3.3 Divergence time estimation

The results of our treePL (Fig. 4; Table S10) and BEAST (Fig. S8; Table S11) analyses show similar divergence time in Cannabaceae; for brevity, only results from the treePL analysis are presented and discussed below. The optimal smoothing value for the treePL analysis (as indicated by the cross-validation tests) was 10^{-7} (Table S12). The stem age of Cannabaceae was estimated to be in the Late Cretaceous (~88.6 million years ago [Ma]; 95% HPD = ~87.8–89.8 Ma). The main lineages of Cannabaceae, *Aphananthe* (~76.9 Ma; 95% HPD = ~74.2–79.1 Ma), *Gironniera* (~68.8 Ma; 95% HPD = ~66.0–70.8 Ma), *Lozanella* (~65.9 Ma; 95% HPD = ~62.8–67.8 Ma), *Celtis* II, and the quadripartite clade (~63.9 Ma; 95% HPD = ~61.1–65.7 Ma), successively diverged during a period of ca. 25 million years (Late Cretaceous to Early Paleocene). Within the quadripartite clade, *Celtis* I originated ~57.9 Ma (95% HPD = ~56.0–59.5 Ma), followed by *Trema* s.l. (~56.0 Ma; 95% HPD = ~54.0–57.6 Ma), and the CHCP subclade (~53.6 Ma; 95% HPD = ~51.6–55.2 Ma).

3.4 Conflict analyses

Our two phyparts analyses showed that nearly all nuclear gene trees (BS > 70%) supported the monophyly of Cannabaceae and the major clades within the family (Figs. 5, 6). The individual nuclear gene trees were also largely concordant with the intergeneric and higher level relationships in the nuclear species and chloroplast trees. However, we found high levels of gene-tree conflict at four deep nodes in the nuclear species tree with mostly low LPP support values (0.96, 0.75, 0.77, and 0.57) and very short subtending branches as measured in coalescent units: the stem node of *Lozanella* (with 6/90 supporting nuclear genes), the stem node of *Celtis* II (with 6/90 supporting nuclear genes), the crown node of the CHCP subclade (with 3/90 supporting nuclear genes), and the crown node of the clade of CHCP + *Trema* s.l. (with no supporting nuclear genes) (Fig. 5). Concerning the phylogenetic position of *Gironniera*, the nuclear topology of *Gironniera* (as sister to a clade comprising *Lozanella*, *Celtis* II, and the quadripartite clade) was supported by 28/90 nuclear gene trees; the chloroplast topology of *Gironniera* (as sister to *Lozanella*) was supported by 4/90 nuclear gene trees. Although the sister relationship between *Celtis* II and the quadripartite clade was consistently supported in both the nuclear species and chloroplast trees, this topology was supported by only 6/90 nuclear genes. Within the quadripartite clade, the sister relationship between *Trema* s.l. and the CHCP subclade in the nuclear species tree was not supported by any nuclear genes (Fig. 5), whereas the sister relationship between *Trema* s.l. and *Cannabis* + *Humulus* in the chloroplast tree was supported by three nuclear genes (Fig. 6).

3.5 Coalescent simulations and network analysis

Coalescent simulations results using guide trees with branch lengths rescaled by a factor of two (Fig. S9) and four (Fig. 7) were largely congruent. Only results with a factor of four are presented and discussed in detail below. We found that the topologies of the chloroplast tree that were in conflict with those of the nuclear species tree were present in a considerable proportion of 1000 simulated trees, indicating the chloroplast topologies were within ILS predictions. For example, 45.2% of the simulated trees supported the chloroplast topology of *Chaetachme* + *Pteroceltis* sister to a clade comprising *Cannabis* + *Humulus* and *Trema* s.l. The sister relationship between *Cannabis* + *Humulus* and *Trema* s.l. in the chloroplast tree was supported by 25.6% of the simulated trees, whereas the sister relationship between *Cannabis* + *Humulus* and *Chaetachme* + *Pteroceltis* (from the nuclear species tree) was supported by 48.1% of the simulated trees. Additionally, 10.2% of the simulated trees supported *Gironniera* as sister to *Lozanella* (the chloroplast topology), while most of the simulated trees (~63.9%) supported *Gironniera* as sister to a clade comprising *Lozanella*, *Celtis* II, and the quadripartite clade (the nuclear species-tree topology).

In PhyloNet analyses based on the "11-taxon dataset," a phylogenetic network with three reticulation events (Fig. 8) was inferred as the best fit model, with the highest log likelihood probability (−877.1451) and the lowest AIC, AICc, and BIC scores (1806.290, 1830.497, and 1869.799, respectively) (Table 2). This best network suggested gene flow between the *Celtis* II clade and the ancestor of the quadripartite clade, with an inheritance probability of 0.11. Within the quadripartite clade, we inferred possible gene flow between an extinct lineage and the ancestor of CHCP + *Trema* s.l., with an inheritance probability of 0.0912. Furthermore, the optimal networks based on the random taxon datasets (i.e., "Random1-11-taxon dataset" and "Random2-11-taxon dataset") all showed the same reticulation scheme with two of the three reticulation events detected using "11-taxon dataset." This suggests relatively consistent PhyloNet results with different taxon sampling in Cannabaceae. All results from PhyloNet analyses (with one to five reticulations) are shown in Figs. S10 and S11.

4 Discussion

4.1 New insights into the phylogeny of Cannabaceae

Our phylogenetic analyses based on many nuclear and chloroplast loci and including broad taxonomic sampling provide new insights into relationships among the main clades of Cannabaceae (Figs. 1, 2). Integrating our phylogenetic results with previous morphological and

Fig. 1. The phylogeny of Cannabaceae inferred from maximum likelihood (ML) analysis in RAxML, based on an optimal partition scheme of the concatenated supermatrix of 90 nuclear genes. Bootstrap support (BS) values (%) from the RAxML analysis, ultrafast bootstrap (UFBoot) values (%) from the IQ-TREE analysis, and local posterior probability (LPP) values from the ASTRAL analysis (using gene trees with nodes of low bootstrap support (BS < 10%) collapsed) are shown (BS/UFBoot/LPP), whereas a hyphen denotes that this relationship is not recovered by the species tree from the ASTRAL analysis (and phylogenetic relationships from the RAxML and IQ-TREE analyses are entirely identical). All nodes received 100% bootstrap support values and 1.00 posterior probabilities unless otherwise indicated.

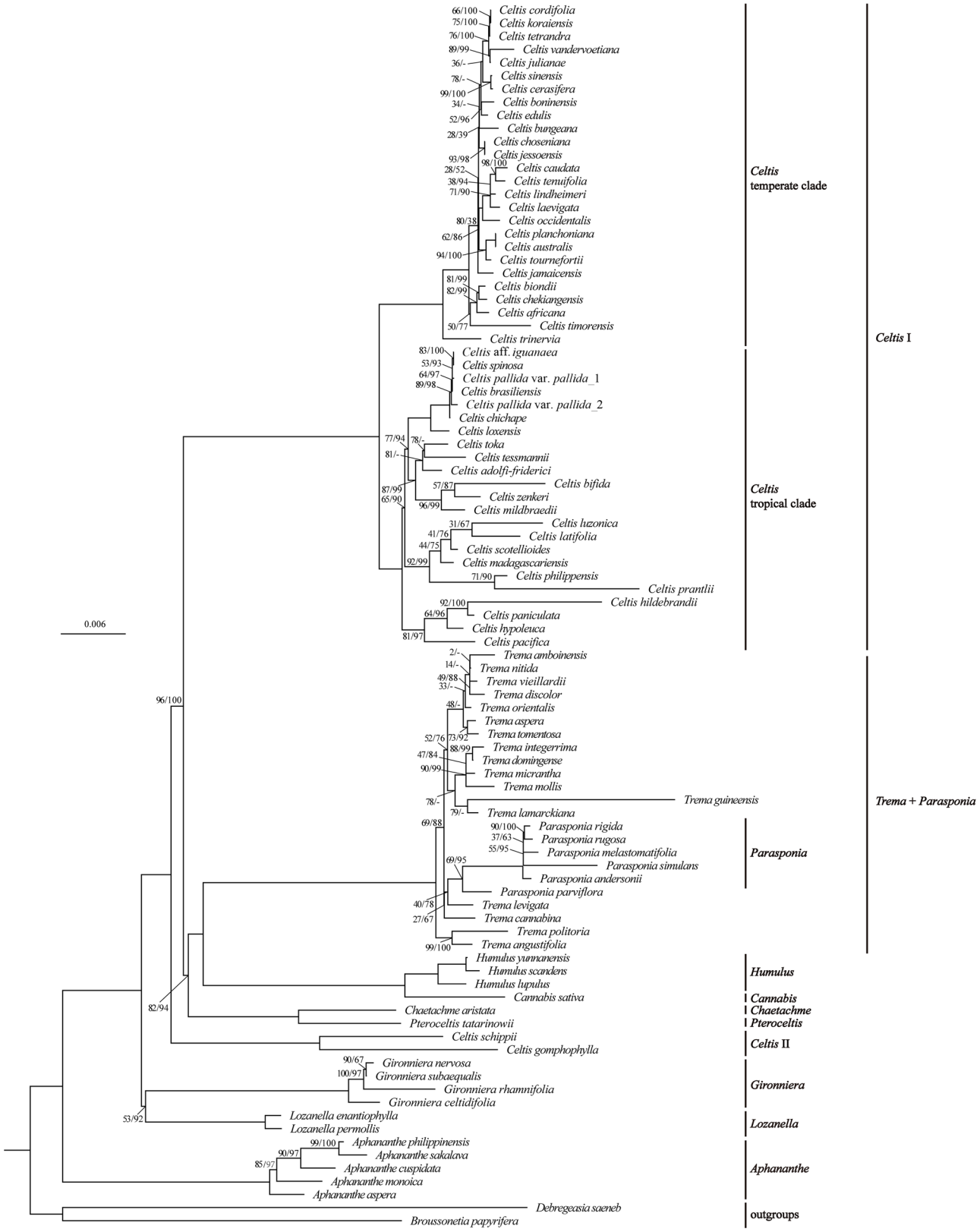

Fig. 2. The phylogeny of Cannabaceae based on the concatenated alignment of 82 chloroplast loci using RAxML. Bootstrap support values (%) from the RAxML analysis and ultrafast bootstrap values (%) from the IQ-TREE analysis are shown above branches (BS/UFBoot), whereas a hyphen denotes that this relationship is not recovered by the IQ-TREE analysis. All nodes received 100% bootstrap support values unless otherwise indicated.

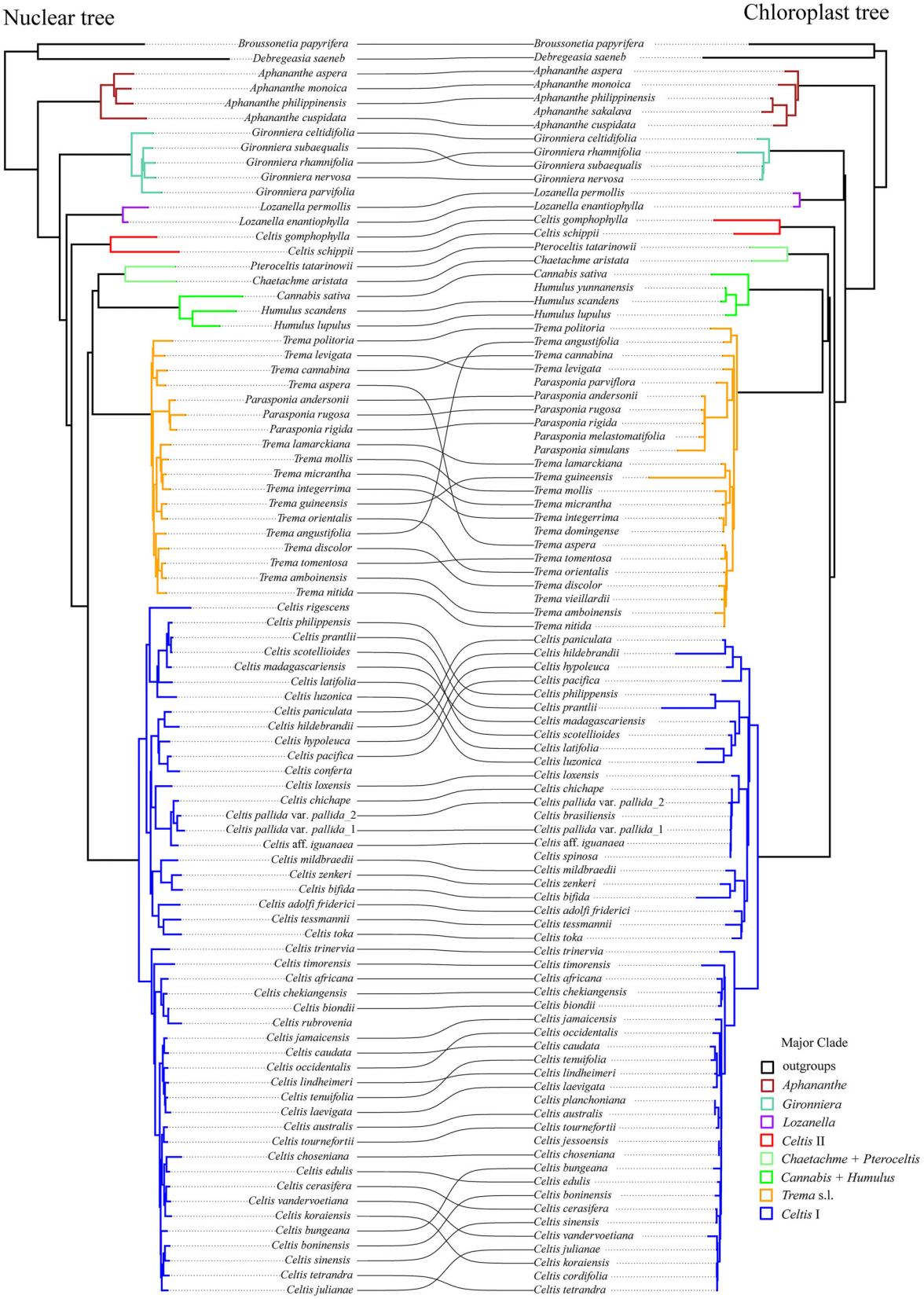

Fig. 3. Tanglegram showing the incongruence between the nuclear (left) and chloroplast (right) trees of Cannabaceae both inferred by partitioned RAxML analyses. Black lines connect the same taxa between the two trees. The colored branches indicate the major clades of Cannabaceae: black, outgroups; brown, *Aphananthe*; medium aquamarine, *Gironniera*; purple, *Lozanella*; red, *Celtis* II; light green, *Chaetachme* + *Pteroceltis*; green, *Cannabis* + *Humulus*; orange, *Trema* s.l.; and blue, *Celtis* I.

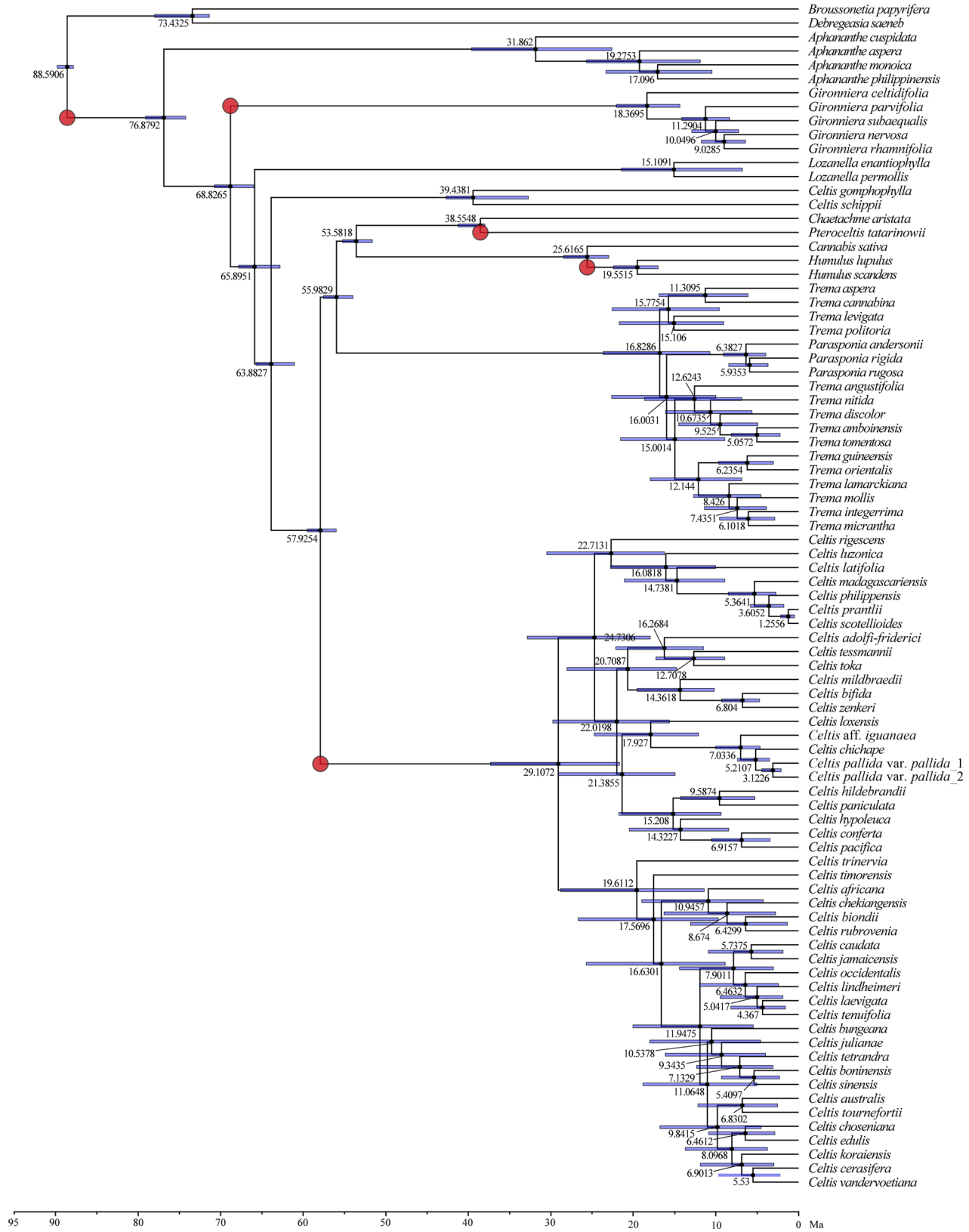

Fig. 4. A time-calibrated tree of Cannabaceae inferred by treePL with five fossil calibrations (red circles). The estimates of the median age and the 95% highest posterior densities (HPD) (blue node bars) (Ma) for each node are shown.

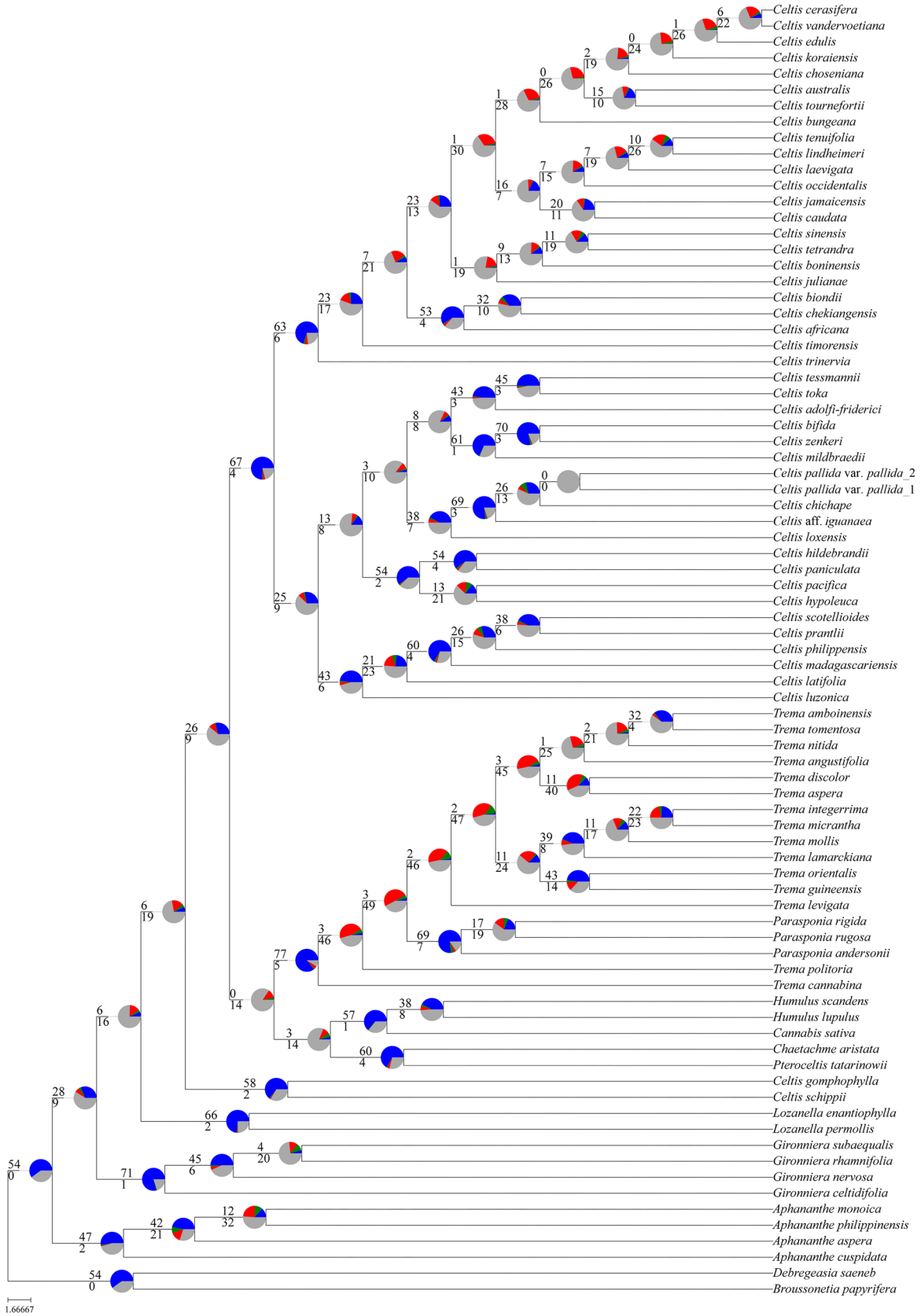

Fig. 5. Continued

molecular studies, we discuss below the evolutionary and morphological significance of our inferred topologies.

4.1.1 *Aphananthe*

Based on morphology and anatomy, Manchester (1989) highlighted that *Aphananthe* appeared to bridge Ulmaceae to Cannabaceae. *Aphananthe* species indeed share some traits with Ulmaceae (i.e., asymmetric ovules and the presence of flavonols), but in terms of pollen structure, leaf vernation pattern, and gynoecial vasculature, they are consistent with other members of Cannabaceae (Yang et al., 2013). Both our nuclear and chloroplast trees fully support *Aphananthe* as sister to all other genera of Cannabaceae (Figs. 1, 2), which is consistent with previous molecular phylogenetic studies (i.e., Sytsma et al., 2002; Sattarian, 2006; Yang et al., 2013; Sun et al., 2016; van Velzen et al., 2018; Zhang et al., 2018b; Jin et al., 2020) and suggests that similarities with Ulmaceae might be symplesiomorphies.

4.1.2 *Gironniera* and *Lozanella*

The phylogenetic positions of *Gironniera* and *Lozanella* have long been controversial. Several previous studies based on chloroplast DNA (Sattarian, 2006; van Velzen et al., 2018) suggested that *Lozanella* and *Gironniera* were successive sisters to the quadripartite clade—that is, (*Lozanella*, (*Gironniera,* the quadripartite clade)). However, Yang et al. (2013) and Sun et al. (2016) argued that *Gironniera* was instead sister to *Lozanella* + the quadripartite clade—that is, (*Gironniera*, (*Lozanella*, the quadripartite clade))—based on several chloroplast and mitochondrial genes. Overall, our results from nuclear phylogenetic analyses showed a slightly different yet fully supported topology, not recovered before, with *Gironniera*, *Lozanella*, and *Celtis* II as successively sister to the quadripartite clade, that is, (*Gironniera*, (*Lozanella*, (*Celtis* II, the quadripartite clade))) (BS = 100%; UFBoot = 100%; LPP = 1.00) (Fig. 1). Nevertheless, our chloroplast-based tree suggested that *Gironniera* and *Lozanella* formed a sister relationship with low support (BS = 53%; UFBoot = 92%) (Fig. 2), consistent with the study of Zhang et al. (2018b) based on plastome data yet with full support (BS = 100%). This suggests that the chloroplast-based relationship of *Gironniera* and *Lozanella* is unlikely to be a product of systematic errors, and the noise in our chloroplast dataset may explain the low support recovered here. *Gironniera* species have alternate leaves and persistent sepals, traits regarded as plesiomorphic for Cannabaceae based on previous ancestral character reconstruction (Yang et al., 2013). In contrast, *Lozanella* species have opposite leaves, which is a morphological autapomorphy for the genus, as suggested by Yang et al. (2013). Moreover, the endocarp body of *Lozanella* is more similar to those of the quadripartite clade (i.e., *Trema* s.l.) than to that found in *Gironniera* (Kravtsova & Wilmot-Dear, 2013). These lines of morphological evidence lend support to our nuclear topology, with *Gironniera* and *Lozanella* successively sister to the remaining genera.

Additionally, phyparts analyses showed that 28/90 informative nuclear genes (i.e., those with BS > 70% for the relevant branch) support the nuclear topology (Fig. 5), whereas only 4/90 nuclear genes support a sister relationship between *Gironniera* and *Lozanella* (i.e., the chloroplast topology, Fig. 6). Our coalescent simulations yielded 10.2% of trees in support of the chloroplast topology (Fig. 7), supporting ILS as an explanation for the observed chloroplast conflict. Furthermore, we did not detect any possible hybridization between *Gironniera* and *Lozanella* in the phylogenetic networks despite their conflicting chloroplast placement (Figs. 8, S11), and to date no current and past geographical overlap is known between *Gironniera* (tropical and subtropical Asia) and *Lozanella* (Mexico to Venezuela and Peru). We therefore argue that ILS might have occurred during the rapid divergence of these two lineages in the late Upper Cretaceous (~68.8 to ~65.9 Ma) (Figs. 4, 7) and that this is the most plausible source of the observed cyto-nuclear discordance.

4.1.3 *Celtis* II clade

Celtis, the most species-rich genus in Cannabaceae, has long been regarded as a monophyletic group (Wiegrefe et al., 1998; Song et al., 2001; Sytsma et al., 2002; Yang et al., 2013; Sun et al., 2016; Jin et al., 2020; Liu et al., 2021). *Celtis* has been reported to include two sister clades corresponding to a tropical evergreen clade and a temperate deciduous clade (Jin et al., 2020), but the study by Jin et al. (2020) only included species from our *Celtis* I clade (i.e., *Celtis gomphophylla* Baker and *Celtis schippii* Standl were not sampled). However, several earlier phylogenetic studies including *C. gomphophylla* and/or *C. schippii* (based on a few molecular loci or phenotypic traits) found that these two species formed a clade (named *Celtis* II here) positioned either sister to the remaining *Celtis* species (i.e., *Celtis* I; Sattarian, 2006; Liu et al., 2021) or sister to the quadripartite clade (including *Celtis* I; Yang et al., 2013), suggesting the possibility of a non-monophyletic *Celtis*. With our expanded sampling of *Celtis* (i.e., 48/73 and 50/73 *Celtis* species in the nuclear and chloroplast datasets, respectively), our nuclear and chloroplast species trees both fully support a sister relationship between *C. gomphophylla* and *C. schippii*, with this clade strongly supported as sister to the quadripartite clade (BS = 99%; UFBoot = 100%; LPP = 0.75 in the nuclear species trees, and BS/UFBoot = 100% in the chloroplast tree) (Figs. 1, 2), thus supporting the non-monophyly of *Celtis*.

Although the sister relationship between *Celtis* II and the quadripartite clade was only supported by 6/90 informative nuclear genes (Fig. 5), the remaining informative genes (19/90) supported various conflicting topologies and no main alternative. The optimal network of Cannabaceae suggested

Fig. 5. Nuclear species tree of Cannabaceae inferred by ASTRAL-III based on the gene trees with nodes of low bootstrap support (BS < 10%) collapsed, showing gene-tree concordance and conflict of 90 nuclear genes based on the phyparts results. Pie charts at each node present the proportion of gene trees in concordance (blue), conflict (green, a common alternative; red, the remaining alternatives), and uninformative (gray) with that bipartition. Numbers above and below the branches indicate the numbers of concordant and conflicting genes at that bipartition, respectively.

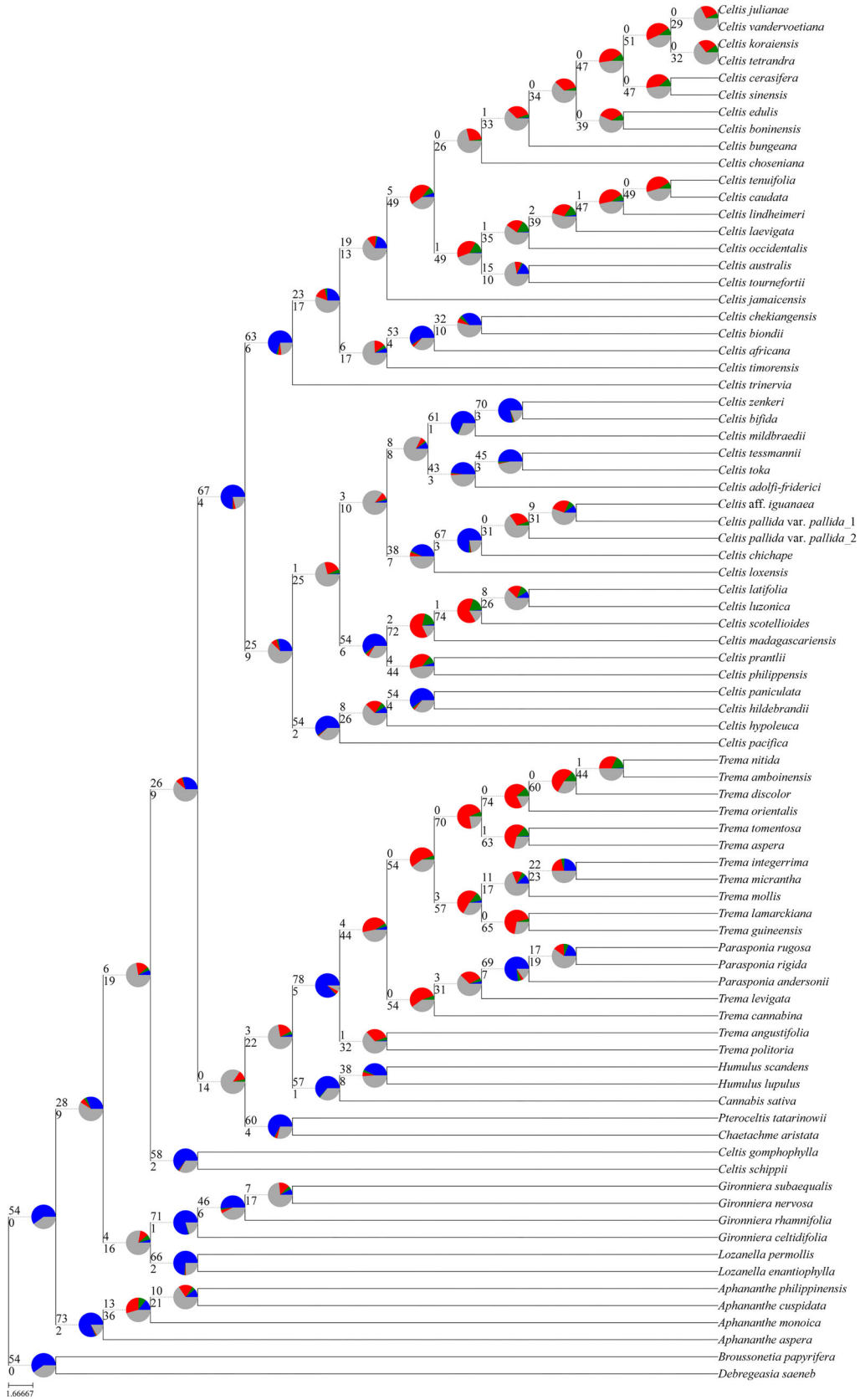

Fig. 6. Continued

possible gene flow (with an inheritance probability of 0.11, indicating asymmetric gene flow) between the ancestor of the quadripartite clade and the *Celtis* II clade (Fig. 8). Accordingly, we argue that the considerable gene-tree conflict observed at the crown node of the clade comprising the *Celtis* II clade and the quadripartite clade may be a product of both ILS and gene flow during the rapid divergence that took place between the *Celtis* II clade and the quadripartite clade in the early Paleocene (~65.9 to ~63.9 Ma) (Figs. 4, 8).

4.1.4 The quadripartite clade

The phylogenetic relationships among the four main lineages of the quadripartite clade (i.e., *Cannabis* + *Humulus*, *Chaetachme* + *Pteroceltis*, *Trema* s.l., and *Celtis* I) were contentious in previous studies (Fig. S1). For instance, Yang et al.

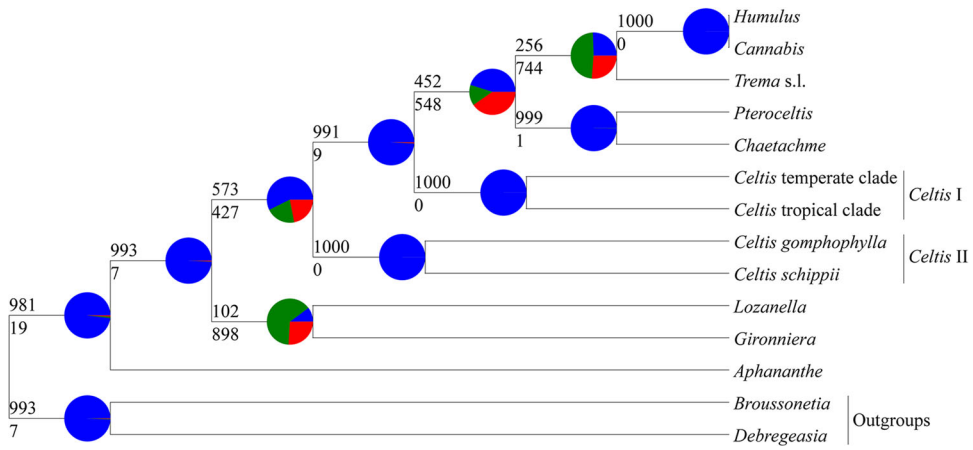

Fig. 7. A summary of 1000 chloroplast trees simulated under the coalescent and using the nuclear ASTRAL species tree (with branch lengths rescaled by four) as the guide tree, obtained by mapping the simulated trees to the chloroplast tree using phyparts. Pie charts at each node denote the proportion of simulated trees in concordance (blue) and conflict (green, a common alternative; red, the remaining alternatives) with that bipartition. Numbers above and below the branches indicate the number of concordant and conflicting simulated trees, respectively.

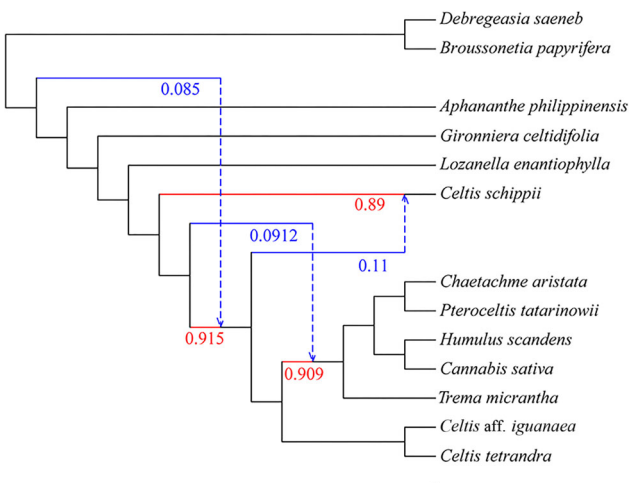

Fig. 8. The optimal phylogenetic network of Cannabaceae inferred by PhyloNet based on the "11-taxon dataset." Hybrid edges were annotated with their inheritance probabilities (γ). The blue edge denotes the minor edge with $\gamma < 0.5$, whereas the red edge denotes the major edge with $\gamma > 0.5$.

Fig. 6. Chloroplast tree of Cannabaceae, showing gene-tree concordance and the conflict of 90 nuclear genes with the chloroplast relationships, based on the phyparts results. Pie charts at each node present the proportion of gene trees in concordance (blue), conflict (green, a common alternative; red, the remaining alternatives), and uninformative (gray) with that bipartition. Numbers above and below the branches indicate the numbers of concordant and conflicting genes at that bipartition, respectively.

Table 2 Model selection among species networks and bifurcating trees of Cannabaceae

Taxon dataset	Trees with 13 representative taxa	Number of reticulations allowed	Ln (likelihood)	Parameters	Loci number	Number of hybridization events detected	Information criteria		
							AIC	AICc	BIC
11-taxon dataset	Chloroplast ML tree	N/A	−952.2916	23	85	N/A	1950.583	1968.682	2006.764
	Nuclear RAxML tree	N/A	−951.2843	23	85	N/A	1948.569	1966.667	2004.750
	Nuclear ASTRAL tree	N/A	−951.2842	23	85	N/A	1948.568	1966.667	2004.749
	Nuclear network 1	1	−904.0122	24	85	1	1856.024	1876.024	1914.648
	Nuclear network 2	2	−891.0121	25	85	2	1832.024	1854.058	1893.090
	Nuclear network 3	3	−890.9994	25	85	2	1831.999	1854.033	1893.065
	Nuclear network 4	**4**	**−877.1451**	**26**	**85**	**3**	**1806.290**	**1830.497**	**1869.799**
	Nuclear network 5	5	−891.0331	25	85	2	1832.066	1854.100	1893.132
Random1-11-taxon dataset	Chloroplast ML tree	N/A	−896.4903	23	82	N/A	1838.981	1858.015	1894.335
	Nuclear RAxML tree	N/A	−915.2528	23	82	N/A	1876.506	1895.540	1931.860
	Nuclear ASTRAL tree	N/A	−915.2527	23	82	N/A	1876.505	1895.540	1931.860
	Nuclear network 1	1	−856.7632	24	82	1	1761.526	1782.579	1819.288
	Nuclear network 2	2	−856.7697	25	82	2	1763.539	1786.754	1823.707
	Nuclear network 3	3	−856.7696	25	82	2	1763.539	1786.753	1823.707
	Nuclear network 4	**4**	**−850.9938**	**25**	**82**	**2**	**1751.988**	**1775.202**	**1812.156**
	Nuclear network 5	5	−856.1876	27	82	4	1766.375	1794.375	1831.357
Random2-11-taxon dataset	Chloroplast ML tree	N/A	−851.3097	23	83	N/A	1748.619	1767.331	1804.253
	Nuclear RAxML tree	N/A	−850.2774	23	83	N/A	1746.555	1765.267	1802.188
	Nuclear ASTRAL tree	N/A	−850.2773	23	83	N/A	1746.555	1765.266	1802.188
	Nuclear network 1	1	−791.5099	24	83	1	1631.020	1651.709	1689.072
	Nuclear network 2	2	−790.6336	25	83	2	1631.267	1654.074	1691.738
	Nuclear network 3	3	−784.0600	26	83	3	1620.120	1645.191	1683.010
	Nuclear network 4	4	−790.3840	25	83	2	1630.768	1653.575	1691.239
	Nuclear network 5	**5**	**−785.8508**	**25**	**83**	**2**	**1621.702**	**1644.509**	**1682.173**

The optimal model with the lowest information criterion is in bold. "N/A" means that it does not apply to the case in question. AIC, Akaike Information Criterion; AICc, corrected Akaike Information Criterion; BIC, Bayesian Information Criterion; ML, maximum likelihood.

(2013) recovered *Cannabis* + *Humulus* as sister to the clade (*Celtis* I, (*Trema* s.l., *Chaetachme* + *Pteroceltis*)) (BS = 100%) based on four chloroplast genes. Sun et al. (2016) moderately supported a sister relationship between the clade (*Celtis* I, *Trema* s.l.) and the clade (*Chaetachme* + *Pteroceltis*, *Cannabis* + *Humulus*) (BS = 82%) using three chloroplast genes and one mitochondrial gene. Jin et al. (2020) fully supported (BS = 100%) *Chaetachme* + *Pteroceltis* as sister to the clade (*Celtis* I, (*Trema* s.l., *Cannabis* + *Humulus*)) using five chloroplast and three nuclear ribosomal loci. Based on extensive sampling of both taxa and loci, all our nuclear trees (based on concatenation and coalescent-based approaches) revealed a topology not recovered previously for the quadripartite clade, with *Celtis* I strongly supported as sister to the remainder of the clade: (*Trema* s.l., (*Chaetachme* + *Pteroceltis*, *Cannabis* + *Humulus*)) (Fig. 1). However, we observed clear cyto-nuclear discordance with respect to relationships among *Trema* s.l., *Chaetachme* + *Pteroceltis*, and *Cannabis* + *Humulus* (Fig. 3). Our chloroplast tree (Fig. 2) placed *Trema* s.l. sister to *Cannabis* + *Humulus*, with these together sister to the clade of *Chaetachme* + *Pteroceltis*—that is, (*Chaetachme* + *Pteroceltis*, (*Trema* s.l., *Cannabis* + *Humulus*))—which is consistent with the results of Zhang et al. (2018b) based on complete plastome. From the perspective of phenotypic data, *Pteroceltis*, *Humulus*, and *Cannabis* have been considered closely related because they all possess sieve tube plastids with starch grains (Behnke, 1989). Coalescent simulations (Fig. 7) showed that the chloroplast topology of *Trema* s.l. sister to *Cannabis* + *Humulus* was frequent in the simulated trees (25.6%), suggesting that this relationship in the chloroplast tree could be a product of ILS. Given that gene flow was not detected between *Trema* s.l. and the *Cannabis* + *Humulus* clade in the PhyloNet analyses (Figs. 8, S11), we suggest that ILS occurred during rapid speciation events in the late Paleocene and early Eocene (i.e., ~57.9 to ~53.6 Ma; Fig. 4) and might be the primary cause of cyto-nuclear discordance concerning the position of *Trema* s.l.

4.2 The merger of *Trema* s.s. and *Parasponia*

Recent molecular phylogenetic studies have consistently indicated that, as originally circumscribed, *Trema* s.s. is paraphyletic, with *Parasponia* embedded within it (Sytsma et al., 2002; Yesson et al., 2004; Sattarian, 2006; Yang et al., 2013; Sun et al., 2016; van Velzen et al., 2018; Jin et al., 2020). Based on molecular and morphological evidence, Yang et al. (2013) suggested that *Parasponia* should be subsumed into *Trema* s.s. However, because the type species (*Parasponia parviflora* Miq.) was not sampled in their study, Yang et al. (2013) refrained from making formal name changes to the species of *Parasponia*. Nevertheless, Christenhusz et al. (2018) formally transferred all species of *Parasponia* to *Trema* in "GLOVAP Nomenclature Part 1, The Global Flora," referencing the study of Yang et al. (2013). This nomenclatural change was not immediately adopted in subsequent studies (i.e., van Velzen et al., 2018, 2019; Bu et al., 2020; Rutten et al., 2020; Shen et al., 2021; Soyano et al., 2021), possibly because the latest taxonomic changes were not widely known. Species of *Trema* s.s. occur across the tropics; those of *Parasponia* are narrowly distributed in Southeast Asia and some Pacific islands. Species of both *Trema* s.s. and *Parasponia* are pioneer plants of early secondary forests or volcanic ash zones (Soepadmo & Lumpur, 1977). Species of *Parasponia* are morphologically very similar to those of *Trema* s.s. regarding leaf shape, secondary venation (pinnate), leaf cystolith structure (pegged), flower sexuality, pollen aperture number (diporate), and embryo shape (curved) (Akkermans et al., 1978; Zavada & Kim, 1996). However, several features distinguish these taxa: the latter has connate intrapetiolar stipules rather than free extrapetiolar stipules (Fig. S12) and imbricate perianth lobes on male flowers rather than induplicate valvate perianth lobes (Fig. S13) (Soepadmo & Lumpur, 1977; Yesson et al., 2004). Perhaps the most well known difference between *Parasponia* and *Trema* s.s. is that *Parasponia* is the only non-legume plant lineage to have evolved a N_2-fixation mutualism with rhizobia, providing a novel system for the study of this ecologically important mutualism (Akkermans et al., 1978; Behm et al., 2014; van Velzen et al., 2018).

In our study, phylogenetic analyses based on both nuclear data (with 3/10 *Parasponia* species and 13/15 *Trema* s.s. sampled) and chloroplast data (with 6/10 *Parasponia* species including the type species and 15/15 species of *Trema* s.s. sampled) all fully support a clade comprising *Parasponia* and *Trema* s.s. (Figs. 1, 2). Our results therefore support the formal merger of *Parasponia* and *Trema* s.s. by Christenhusz et al. (2018). We suggest that *Parasponia* should be treated as a synonym of *Trema* s.l. and that the species formerly treated as *Parasponia* should collectively be referred to as "the *Parasponia* clade" in future studies. Current names for the species of the *Parasponia* clade are listed in Table 3.

4.3 Reinstitution of *Sparrea* Hunziker & Dottori as a phylogenetically and morphologically distinct genus in Cannabaceae

Our phylogenies of Cannabaceae based on both nuclear and chloroplast data show that *Celtis*, as currently circumscribed, is non-monophyletic, with two distinct clades: (A) *Celtis* I (including most *Celtis* species), sister to the rest of the quadripartite clade; and (B) *Celtis* II (including *C. gomphophylla* and *C. schippii*), sister to the entire quadripartite clade (Figs. 1, 2). *Celtis* I species are commonly monoecious trees with serrate leaf margins, caducous sepals, one short style (with 2-lobed stigma) on mature fruit, and rough and light color endocarps. In contrast to species of *Celtis* I, those of *Celtis* II commonly have entire leaf margins, persistent sepals on the mature fruit, and smooth and dark brown/black endocarps. *Celtis gomphophylla*, in particular, is usually a dioecious deciduous tree with unarmed branch and nearly entire leaf margin, persistent sepals, two styles (with unlobed stigma) on the mature fruit (as shown by the following specimen: https://science.mnhn.fr; herbarium collection number: P00562034) and brown endocarps, with distribution in tropical and southern Africa, Comoros, and Madagascar (Rendle, 1916; Leroy, 1952; Polhill, 1966; Sattarian, 2006; https://powo.science.kew.org/taxon/urn:lsid:ipni.org:names:851066-1). *Celtis schippii* occurs in the Neotropics (specifically, central and western South America) but differs from other *Celtis* species in the region in having unarmed branches, persistent sepals on the mature fruit, and membranaceous and black endocarps (Berg & Dahlberg, 2001; Zamengo, 2019; Zamengo et al., 2020). In 1978, Argentinian scientists documented the remarkable morphological differences between the embryos of *C. schippii*

Table 3 Revised names of species in the *Parasponia* clade according to Christenhusz et al. (2018)

Combinatio nova	Basionym	Original publication
Trema andersonii (Planch.) Byng & Christenh.	*Parasponia andersonii* Planch.	Prodr. 17: 195 1873.
Trema aspera (Blume) Byng & Christenh.	*Parasponia aspera* Blume	Mus. Bot. 2: 66 1856.
Trema eurhyncha (Miq.) Byng & Christenh.	*Parasponia eurhyncha* Miq.	Fl. Ned. Ind., Eerste Bijv. 411 1861.
Trema melastomatifolia (J.J.Sm.) Byng & Christenh.	*Parasponia melastomatifolia* J.J.Sm.	Nova Guinea 8: 891 1912.
Trema parviflora (Miq.) Byng & Christenh.	*Parasponia parviflora* Miq.	Pl. Jungh. 69 1851.
Trema paucinervia (Merr. & L.M.Perry) Byng & Christenh.	*Parasponia paucinervia* Merr. & L.M.Perry	J. Arnold Arbor. 20: 324 1939.
Trema rigida (Merr. & L.M.Perry) Byng & Christenh.	*Parasponia rigida* Merr. & L.M.Perry	J. Arnold Arbor. 22: 254 1941.
Trema rugosa (Blume) Byng & Christenh.	*Parasponia rugosa* Blume	Mus. Bot. 2: 66 1856.
Trema similis (Blume) Byng & Christenh.	*Parasponia similis* Blume	Mus. Bot. 2: 66 1856.
Trema simulans (Merr. & L.M.Perry) Byng & Christenh.	*Parasponia simulans* Merr. & L.M.Perry	J. Arnold Arbor. 22: 255 1941.

and those of other extant species and genera of Cannabaceae and Ulmaceae, prompting them to name a new genus, *Sparrea* Hunziker & Dottori, for this unique species (Hunziker & Dottori, 1978). In light of this morphological evidence and our phylogenetic results, we suggest that the generic name *Sparrea* should be applied to both species of the *Celtis* II clade.

5 Conclusions

This work presents the most comprehensive phylogenetic framework for Cannabaceae to date and offers new insights into previously controversial relationships among several major lineages. However, we also observed deep cyto-nuclear discordances concerning the positions of *Lozanella* and *Trema* s.l. and high levels of gene-tree conflict at several deep nodes in Cannabaceae. Our coalescent simulations and network analyses suggested that ILS is the more plausible source of the observed cyto-nuclear discordances. At the same time, a combination of both ILS and gene flow might explain the gene-tree heterogeneity associated with the rapid diversification of clades recognized at the generic level. However, our genomic sampling (including 90 nuclear genes) was insufficient to detect past gene flow with certainty in PhyloNet analyses. Future studies leveraging more complete nuclear genomic datasets will be necessary to evaluate more fully the presence and/or extent of past hybridization events in Cannabaceae. In light of both morphological data and our phylogenetic results, we support the recent merger of *Parasponia* into *Trema* s.l.; we also call for the reinstatement of *Sparrea* Hunziker & Dottori (i.e., *Celtis* II) as a distinct genus and expand its circumscription to include a second species: *Sparrea gomphophylla* (Baker) X.G.Fu & T.S.Yi, comb. nov. Basionym: *Celtis gomphophylla* Baker, J. Linn. Soc., Bot. 22: 521. 1887. These changes form the basis of a revised generic classification of Cannabaceae comprising 10 monophyletic genera: *Aphananthe*, *Cannabis*, *Celtis*, *Chaetachme*, *Gironniera*, *Humulus*, *Lozanella*, *Pteroceltis*, *Sparrea*, and *Trema*. The phylogenetic framework and revised classification of Cannabaceae presented here should offer a solid foundation for future evolutionary studies of the family.

Acknowledgements

This research was supported by the National Natural Science Foundation of China, key international (regional) cooperative research project (No. 31720103903), the Strategic Priority Research Program of Chinese Academy of Sciences (No. XDB31000000), the Science and Technology Basic Resources Investigation Program of China (No. 2019FY100900), the Large-scale Scientific Facilities of the Chinese Academy of Sciences (No. 2017-LSF-GBOWS-02), the National Natural Science Foundation of China (No. 31270274), the Yunling International High-end Experts Program of Yunnan Province, China (No. YNQR-GDWG-2017-002 and No. YNQR-GDWG-2018-012), the CAS President's International Fellowship Initiative (No. 2020PB0009), the China Postdoctoral Science Foundation (CPSF) International Postdoctoral Exchange Program, and the CAS Special Research Assistant Project. This work was also supported in part by USA Department of Energy grant DE-SC0018247. We are grateful to the following institutes for providing specimens or silica-dried materials: Herbarium of Kunming Institute of Botany, Chinese Academy of Sciences; the Germplasm Bank of Wild Species in the Southwest China and Molecular Biology Experiment Center, Kunming Institute of Botany, Chinese Academy of Sciences; the herbarium of the California Academy of Sciences; the Harvard University Herbaria; University of Texas at Austin Herbarium; The Ohio State University Herbarium; National Herbarium of the Netherlands; and the herbaria of the Royal Botanic Gardens, Kew, the Missouri Botanical Garden, the New York Botanical Garden, the San Francisco Botanical Garden, and the California Botanical Garden. We are also grateful to Prof. Susanne S. Renner for providing Cannabaceae materials; to Jiajin Wu for help with sampling; to Prof. Huafeng Wang, Dr. Diego F. Morales-Briones, Dr. Nelson Zamora Villalobos, Dr. Rong Zhang, Dr. Hui Liu, Siyun Chen, Chenxuan Yang, Yingying Yang, and Henrique Borges Zamengo for their generous technical support and necessary assistance; and to the iFlora High Performance Computing Center of Germplasm Bank of Wild Species (iFlora HPC Center of GBOWS, KIB, CAS) for computing.

References

Akkermans ADL, Abdulkadir S, Trinick MJ. 1978. Nitrogen-fixing root nodules in Ulmaceae. *Nature* 274: 190.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19: 455–477.

Behm JE, Geurts R, Kiers ET. 2014. *Parasponia*: A novel system for studying mutualism stability. *Trends in Plant Science* 19: 757–763.

Behnke HD. 1989. Sieve-element plastids, phloem proteins, and the evolution of flowering plants. IV. Hamamelidae. In: Crane PR, Blackmore S eds. *Evolution, systematics, and fossil history of the hamamelidae*. Oxford: Clarendon Press. 105–128.

Behre KE. 1999. The history of beer additives in Europe—A review. *Vegetation History and Archaeobotany* 8: 35–48.

Berg CC, Dahlberg SV. 2001. A revision of *Celtis* subg. *Mertensia* (Ulmaceae). *Brittonia* 53: 66–81.

Biendl M, Pinzl C. 2007. *Arzneipflanze Hopfen: Anwendungen, Wirkungen, Geschichte*. Wolnzach: Deutsches Hopfenmuseum. 17–38.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120.

Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: A software platform for Bayesian evolutionary analysis. *PloS Computational Biology* 10: e1003537.

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

Bu F-J, Rutten L, Roswanjaya YP, Kulikova O, Rodriguez-Franco M, Ott T, Bisseling T, van Zeijl A, Geurts R. 2020. Mutant analysis in the nonlegume *parasponia andersonii* identifies nin and nf-ya1 transcription factors as a core genetic network in nitrogen-fixing nodule symbioses. *New Phytologist* 226: 541–554.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: Architecture and applications. *BMC Bioinformatics* 10: 421.

Cao Z, Liu X-H, Ogilvie HA, Yan Z, Nakhleh L. 2019. Practical aspects of phylogenetic network analysis using PhyloNet. *bioRxiv*. https://doi.org/10.1101/746362

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25: 1972–1973.

Christenhusz MJM, Fay MF, Byng JW. 2018. GLOVAP nomenclature part 1. In: Christenhusz MJM ed. *The global flora: A practical flora to vascular plant species of the world*. Bradford: Plant Gateway Ltd. 1–155.

Doyle JJ. 1992. Gene trees and species trees: Molecular systematics as one-character taxonomy. *Systematic Botany* 19: 144–163.

Doyle JJ. 2022. Defining coalescent genes: Theory meets practice in organelle phylogenomics. *Systematic Biology* 71: 476–489.

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

Engler A, Prantl K. 1893. *Die natürlichen Pflanzenfamilien*. Berlin: Bornträger. 3: 202–230.

Folk RA, Kates HR, LaFrance R, Soltis DE, Soltis PS, Guralnick RP. 2021. High-throughput methods for efficiently building massive phylogenies from natural history collections. *Applications in Plant Sciences* 9: e11410.

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

García N, Folk RA, Meerow AW, Chamala S, Gitzendanner MS, de Oliveira RS, Soltis DE, Soltis PS. 2017. Deep reticulation and incomplete lineage sorting obscure the diploid phylogeny of rain-lilies and allies (Amaryllidaceae tribe Hippeastreae). *Molecular Phylogenetics and Evolution* 111: 231–247.

Gitzendanner MA, Soltis PS, Wong GKS, Ruhfel BR, Soltis DE. 2018. Plastid phylogenomic analysis of green plants: A billion years of evolutionary history. *American Journal of Botany* 105: 291–301.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of phyml 3.0. *Systematic Biology* 59: 307–321.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution* 35: 518–522.

Hunziker AT, Dottori NM. 1978. *Sparrea*, nuevo genero de Ulmaceae. *Kurtziana* 11: 25–40.

Jin J-J, Yang M-Q, Fritsch PW, van Velzen R, Li D-Z, Yi T-S. 2020. Born migrators: Historical biogeography of the cosmopolitan family Cannabaceae. *Journal of Systematics and Evolution* 58: 461–473.

Johnson MG, Gardner EM, Liu Y, Medina R, Goffinet B, Shaw AJ, Zerega NJC, Wickett NJ. 2016. HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences* 4: 1600016.

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

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

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012. Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28: 1647–1649.

Kovalchuk I, Pellino M, Rigault P, van Velzen R, Ebersbach J, Ashnest JR, Mau M, Schranz ME, Alcorn J, Laprairie RB, McKay JK, Burbridge C, Schneider D, Vergara D, Kane NC, Sharbel TF. 2020. The genomics of *Cannabis* and its close relatives. *Annual Review of Plant Biology* 71: 713–739.

Kravtsova TI, Wilmot-Dear CM. 2013. Fruit structure in *Lozanella enantiophylla* and *L. permollis* (Celtidaceae). *Botanicheskiĭ Zhurnal* 98: 468–480.

Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. 2017. PartitionFinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution* 34: 772–773.

Leroy JF. 1952. *Ulmacées*. In: Humbert H, Leroy JF eds. *Flore de Madagascar et des Comores: plantes vasculaires*. Tananarive: Imprimerie officielle. 1–15.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760.

Li X-H, Shao J-W, Lu C, Zhang X-P, Qiu Y-X. 2012. Chloroplast phylogeography of a temperate tree *Pteroceltis tatarinowii* (Ulmaceae) in China. *Journal of Systematics and Evolution* 50: 325–333.

Link DHF. 1831. *Handbuch zur Erkennung der nutzbarsten und am häufigsten vorkommenden Gewächse*. Berlin: Spenerschen Buchhandlung. 1–546.

Liu L-X, Zhang Y-H, Li P. 2021. Development of genomic resources for the genus *Celtis* (Cannabaceae) based on genome skimming data. *Plant Diversity* 43: 43–53.

Manchester SR. 1989. Systematics and fossil history of the Ulmaceae. In: Crane PR, Blackmore S eds. *Evolution, systematics and fossil history of the Hamamelidae, Vol. 2: "Higher" Hamamelidae*. Oxford: Clarendon Press. 221–251.

Maurin KJL. 2020. An empirical guide for producing a dated phylogeny with treePL in a maximum likelihood framework. *arXiv*. 2008.07054.

McCauley DE. 1994. Contrasting the distribution of chloroplast DNA and allozyme polymorphism among local populations of Silene alba: Implications for studies of gene flow in plants. *Proceedings of the National Academy of Sciences USA* 91: 8127–8131.

Mendes FK, Hahn MW. 2016. Gene tree discordance causes apparent substitution rate variation. *Systematic Biology* 65: 711–721.

Morales-Briones DF, Kadereit G. 2022. Exploring the possible role of hybridization in the evolution of photosynthetic pathways in *Flaveria* (Asteraceae), the prime model of C4 photosynthesis evolution. *bioRxiv*. https://doi.org/10.1101/2022.01.31.478436

Morales-Briones DF, Kadereit G, Tefarikis DT, Moore MJ, Smith SA, Brockington SF, Timoneda A, Yim WC, Cushman JC, Yang Y. 2021. Disentangling sources of gene tree discordance in phylogenomic data sets: Testing ancient hybridizations in Amaranthaceae s.l. *Systematic Biology* 70: 219–235.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* 32: 268–274.

Polhill RM. 1966. Ulmaceae. In: Hubbard CE, Milne-Redhead E eds. *Flora of tropical East Africa*. London: Crown Agents for the Colonies. 1–15

Pringle H. 1997. Ice age communities may be earliest known net hunters. *Science* 277: 1203–1204.

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology* 67: 901–904.

Ren G-P, Zhang X, Li Y, Ridout K, Serrano-Serrano ML, Yang Y-Z, Liu A, Ravikanth G, Nawaz MA, Mumtaz AS, Salamin N, Fumagalli L. 2021. Large-scale whole-genome resequencing unravels the domestication history of *Cannabis sativa*. *Science Advances* 7: eabg2286.

Rendle AB. 1916. Ulmaceae. In: Oliver D ed. *Flora of tropical Africa*. London: Lovell Reeve & co., Limited. 1–14.

Rendle AB. 1925. *The classification of flowering plants: Dicotyledons*. London: Cambridge University Press. 43–48.

Rutten L, Miyata K, Roswanjaya YP, Huisman R, Bu F-J, Hartog M, Linders S, van Velzen R, van Zeijl A, Bisseling T, Kohlen W, Geurts R. 2020. Duplication of symbiotic lysin motif receptors predates the evolution of nitrogen-fixing nodule symbiosis. *Plant Physiology* 184: 1004–1023.

Sattarian A. 2006. *Contribution to the biosystematics of* Celtis *L. (Celtidaceae) with special emphasis on the African species*. Ph.D. Dissertation. Wageningen: Wageningen University.

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

Shen D-F, Holmer R, Kulikova O, Mannapperuma C, Street NR, Yan Z-C, van der Maden T, Bu F-J, Zhang Y-Y, Geurts R, Magne K. 2021. The BOP-type co-transcriptional regulator NODULE ROOT1 promotes stem secondary growth of the tropical Cannabaceae tree *Parasponia andersonii*. *The Plant Journal* 106: 1366–1386.

Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6: 31.

Smith SA, Brown JW, Walker JF. 2018. So many genes, so little time: A practical approach to divergence-time estimation in the genomic era. *PLoS One* 13: e0197433.

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

Smith SA, O'Meara BC. 2012. treePL: Divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* 28: 2689–2690.

Soepadmo E, Lumpur K. 1977. Ulmaceae. In: Steenis CGGJv, Steenis-Kruseman MJv, Indonesia Departemen Pertanian, Kebun Raya Indonesia, Lembaga Ilmu Pengetahuan Indonesia eds. *Flora Malesiana*. Djakarta: Noordhoff-Kolff. 31–76.

Soltis DE, Smith SA, Cellinese N, Wurdack KJ, Tank DC, Brockington SF, Refulio-Rodriguez NF, Walker JB, Moore MJ, Carlsward BS, Bell CD, Latvis M, Crawley S, Black C, Diouf D, Xi Z-X, Rushworth CA, Gitzendanner MA, Sytsma KJ, Qiu Y-L, Hilu KW, Davis CC, Sanderson MJ, Beaman RS, Olmstead RG, Judd WS, Donoghue MJ, Soltis PS. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. *American Journal of Botany* 98: 704–730.

Song B-H, Li F-Z. 2002. The utility of *trnK* intron 5 region in phylogenetic analysis of Ulmaceae s. l. *Journal of Systematics and Evolution* 40: 125–132.

Song B-H, Wang X-Q, Li F-Z, Hong D-Y. 2001. Further evidence for paraphyly of the Celtidaceae from the chloroplast gene *matK*. *Plant Systematics and Evolution* 228: 107–115.

Soyano T, Liu M, Kawaguchi M, Hayashi M. 2021. Leguminous nodule symbiosis involves recruitment of factors contributing to lateral root development. *Current Opinion in Plant Biology* 59: 102000.

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

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

Stull GW, Soltis PS, Soltis DE, Gitzendanner MA, Smith SA. 2020. Nuclear phylogenomic analyses of asterids conflict with plastome trees and support novel relationships among major lineages. *American Journal of Botany* 107: 790–805.

Sukumaran J, Holder MT. 2010. DendroPy: A Python library for phylogenetic computing. *Bioinformatics* 26: 1569–1571.

Sun M, Naeem R, Su J-X, Cao Z-Y, Burleigh JG, Soltis PS, Soltis DE, Chen Z-D. 2016. Phylogeny of the *Rosidae*: A dense taxon sampling analysis. *Journal of Systematics and Evolution* 54: 363–391.

Sun Q-H, Morales-Briones DF, Wang H-X, Landis JB, Wen J, Wang H-F. 2022. Phylogenomic analyses of the East Asian endemic *Abelia* (Caprifoliaceae) shed insights into the temporal and spatial diversification history with widespread hybridization. *Annals of Botany* 129: 201–216.

Sytsma KJ, Morawetz J, Pires JC, Nepokroeff M, Conti E, Zjhra M, Hall JC, Chase MW. 2002. Urticalean rosids: Circumscription, rosid

ancestry, and phylogenetics based on *rbcL*, *trnL-F*, and *ndhF* sequences. *American Journal of Botany* 89: 1531–1546.

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

van Velzen R, Doyle JJ, Geurts R. 2019. A resurrected scenario: Single gain and massive loss of nitrogen-fixing nodulation. *Trends in Plant Science* 24: 49–57.

van Velzen R, Holmer R, Bu F-J, Rutten L, van Zeijl A, Liu W, Santuari L, Cao Q-Q, Sharma T, Shen D-F, Roswanjaya Y, Wardhani TAK, Kalhor MS, Jansen J, van den Hoogen J, Güngör B, Hartog M, Hontelez J, Verver J, Yang W-C, Schijlen E, Repin R, Schilthuizen M, Schranz ME, Heidstra R, Miyata K, Fedorova E, Kohlen W, Bisseling T, Smit S, Geurts R. 2018. Comparative genomics of the nonlegume *Parasponia* reveals insights into evolution of nitrogen-fixing rhizobium symbioses. *Proceedings of the National Academy of Sciences USA* 115: E4700–E4709.

Walker JF, Walker-Hale N, Vargas OM, Larson DA, Stull GW. 2019. Characterizing gene tree conflict in plastome-inferred phylogenies. *PeerJ* 7: e7747.

Wang H-X, Morales-Briones DF, Moore MJ, Wen J, Wang H-F. 2021. A phylogenomic perspective on gene tree conflict and character evolution in Caprifoliaceae using target enrichment data, with Zabelioideae recognized as a new subfamily. *Journal of Systematics and Evolution* 59: 897–914.

Wen D-Q, Yu Y, Zhu J-F, Nakhleh L. 2018. Inferring phylogenetic networks using PhyloNet. *Systematic Biology* 67: 735–740.

Wiegrefe SJ, Sytsma KJ, Guries RP. 1998. The Ulmaceae, one family or two? Evidence from chloroplast DNA restriction site mapping. *Plant Systematics and Evolution* 210: 249–270.

Yang M-Q, van Velzen R, Bakker FT, Sattarian A, Li D-Z, Yi T-S. 2013. Molecular phylogenetics and character evolution of Cannabaceae. *Taxon* 62: 473–485.

Yesson C, Russell SJ, Parrish T, Dalling JW, Garwood NC. 2004. Phylogenetic framework for *Trema* (Celtidaceae). *Plant Systematics and Evolution* 248: 85–109.

Yu Y, Degnan JH, Nakhleh L. 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genetics* 8: e1002660.

Zamengo HB. 2019. Celtis *L. (Cannabaceae) do Brasil*. Master Dissertation. São Paulo: Instituto de Botânica.

Zamengo HB, Gaglioti AL, Chamorro D, Mogni V, Oakley L, Prado D, Torres RB, de Mattos L, Da-Silva PR, Romaniuc-Neto S. 2020. Nomenclatural novelties in *Celtis* (Cannabaceae) and a preliminary phylogeny of the genus with emphasis on the South American species. *Brazilian Journal of Botany* 43: 947–960.

Zanoli P, Zavatti M. 2008. Pharmacognostic and pharmacological profile of *Humulus lupulus* L. *Journal of Ethnopharmacology* 116: 383–396.

Zavada MS, Kim M. 1996. Phylogenetic analysis of Ulmaceae. *Plant Systematics and Evolution* 200: 13–20.

Zhang C, Sayyari E, Mirarab S. 2017. ASTRAL-III: Increased scalability and impacts of contracting low support branches. In: Meidanis J, Nakhleh L eds. *Comparative genomics*. Cham: Springer. 53–75.

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

Zhang H-L, Jin J-J, Moore MJ, Yi T-S, Li D-Z. 2018b. Plastome characteristics of Cannabaceae. *Plant Diversity* 40: 127–137.

Zhang R, Wang Y-H, Jin J-J, Stull GW, Bruneau A, Cardoso D, De Queiroz LP, Moore MJ, Zhang S-D, Chen S-Y, Wang J, Li D-Z, Yi T-S. 2020. Exploration of plastid phylogenomic conflict yields new insights into the deep relationships of Leguminosae. *Systematic Biology* 69: 613–622.

Zhang S-D, Soltis DE, Yang Y, Li D-Z, Yi T-S. 2011. Multi-gene analysis provides a well-supported phylogeny of Rosales. *Molecular Phylogenetics and Evolution* 60: 21–28.

Zlas J, Stark H, Seligman J, Levy R, Werker E, Breuer A, Mechoulam R. 1993. Early medical use of *Cannabis*. *Nature* 363: 215.

Supplementary Material

The following supplementary material is available online for this article at http://onlinelibrary.wiley.com/doi/10.1111/jse.12920/suppinfo:

Fig. S1. Schematic diagrams of the phylogenetic relationships among genera of Cannabaceae recovered by previous molecular studies.

Fig. S2. Heatmaps showing the recovery efficiency for (A) 100 nuclear genes and (B) 82 chloroplast loci in Hyb-Seq data, assembled by HybPiper. Each row denotes a sample, and each column denotes a gene.

Fig. S3. The species tree inferred by ASTRAL-III based on the gene trees with nodes of low bootstrap support (BS < 10%) collapsed. The local posterior probability (LPP) values are shown on nodes. All nodes received 1.00 LPP values unless otherwise indicated.

Fig. S4. Tanglegram showing incongruence between the nuclear concatenated maximum likelihood (ML) tree (left) inferred by partitioned RAxML analyses and the nuclear ASTRAL species tree (right).

Fig. S5. Tanglegram showing the concordance between two ASTRAL species trees, based on nuclear gene trees that are directly from RAxML analyses (right), and with nodes of low bootstrap support (BS < 10%) collapsed (left).

Fig. S6. Tanglegram showing the incongruence between two concatenated ML trees, based on 90 nuclear genes (left) with more missing data and 49 nuclear genes (right) that all have more than 90% of the sampled species. Both concatenated ML trees were inferred by RAxML under the partitioned GTR-GAMMA model.

Fig. S7. Tanglegram showing the concordance between two concatenated ML trees inferred by RAxML, under a partitioned (left) GTR-GAMMA model and an unpartitioned (right) GTR-GAMMA model. Both concatenated ML trees were inferred from the concatenated alignment of 90 nuclear genes.

Fig. S8. A time-calibrated tree of Cannabaceae inferred by BEAST with five fossil calibrations (red circles). The estimates of the median age and the 95% highest posterior densities (HPD) (blue node bars) (Ma) for each node are shown.

Fig. S9. A summary of 1,000 chloroplast trees simulated under the coalescent model and using the nuclear ASTRAL species tree (with branch lengths rescaled by two) as the guide tree, obtained by mapping the simulated trees to the chloroplast tree using phyparts.

Fig. S10. Phylogenetic networks of Cannabaceae inferred by PhyloNet based on the "11-taxon dataset."

Fig. S11. Phylogenetic networks of Cannabaceae inferred by PhyloNet based on another two different taxon datasets ("Random1-11-taxon dataset" and "Random2-11-taxon dataset").

Fig. S12. Stipule photographs of *Trema* s.s. and *Parasponia*.

Fig. S13. Male flower photographs of *Trema* s.s. and *Parasponia*.
Table S1. Main classification history of Cannabaceae.
Table S2. Sampling information of 83 Cannabaceae species and two outgroup species used in Hyb-Seq.
Table S3. The statistic information of nuclear gene alignments with ambiguously aligned regions removed by trimAl.
Table S4. Sampling information for 24 complete chloroplast genomes of Cannabaceae obtained from GenBank.
Table S5. Accession numbers of the chloroplast DNA loci of Cannabaceae obtained from GenBank.
Table S6. The statistic information of 82 chloroplast DNA loci used for phylogenetic analysis.
Table S7. Five fossil calibrations used for the molecular dating of Cannabaceae.
Table S8. The summary of HybPiper assembly results for nuclear genes.
Table S9. The summary of HybPiper assembly results for chloroplast DNA loci.
Table S10. Estimated divergence times of Cannabaceae and its major clades using treePL.
Table S11. Estimated divergence times of Cannabaceae and its major clades using BEAST.
Table S12. The result of 10 repeated cross-validation analyses for smoothing value in treePL.