MATERIALS AND METHODS

*Sampling Strategy and DNA Extraction*

We sampled 187 species of 165 genera representing 35 of the 36 tribes of Leguminosae recognized by Lewis et al. (2005), as well as major lineages of all six newly defined subfamilies of Leguminosae (LPWG 2013, 2017). To include a representative of the missing tribe Mimozygantheae, the sequences of *matK* (AY944556) and *trnL*-*trnF* (AY944539) of *Mimozyganthus carinatus* from GenBank were included in phylogenetic reconstructions. Eight taxa from five genera of the other three families of Fabales (Polygalaceae, Quillajaceae and Surianaceae) were included as outgroups. Plastomes of 151 species were newly sequenced, and all other publicly available legume plastomes were downloaded from GenBank. Complete plastome sequences may be downloaded from NCBI, and nearly-complete annotated plastome sequences could be found from Dryad (<https://doi.org/10.5061/dryad.1vhhmgqpb>). All accessions are listed in Supplementary Table S1.

Multiple types of materials including fresh leaves, silica-gel dried leaves, and herbarium samples were used for plastome sequencing. Total genomic DNA (gDNA) was isolated from fresh leaves and silica-gel dried leaves using a modified Cetyl Trimethyl Ammonium Bromide (CTAB) method (Doyle and Doyle 1987), using 4% CTAB with approximately 1% polyvinyl polypyrrolidone and 0.2% DL-dithiothreitol (Yang et al. 2014). Genomic DNA was extracted from herbarium samples using the DNeasy Plant Mini Kit of the Tiangen (Tiangen Biotech Co., LTD., Beijing, China). Information on the materials used is presented in Supplementary Table S1.

*Genome Sequencing*

Two different methods were used to obtain plastomes. A total of 31 plastomes were amplified via long-range PCR, employing fifteen primer pairs designed by Zhang et al. (2016). PCR amplicons for each taxon were mixed in roughly equimolar concentrations, and a total of 6 μg of this mixture was used for Illumina Nextera XT library construction (Illumina, San Diego, CA, USA) following the manufacturer’s instructions. Libraries were size-selected for 500 bp inserts. Sequencing with 2 × 150 bp paired-end reads was performed on the Illumina MiSeq at the Plant Germplasm and Genomics Center, Kunming Institute of Botany, Chinese Academy of Sciences (KIB, CAS). For the remaining 120 species, gDNA was fragmented and libraries were size selected for 350 bp inserts. Sequencing with 2 × 150 bp paired-end reads was conducted using an Illumina HiSeq X-Ten at Novogene (Tianjin, China) or an Illumina HiSeq 4000/X-Ten at BGI (Shenzhen, China).

*Plastome Assembly and Annotation*

All raw reads from the plastomes generated by long-range PCR enrichment were quality filtered using the NGS QC Tool Kit (Patel and Jain 2012) with default parameters. The resulting high-quality, paired-end reads were *de novo* assembled into contigs using CLC Genomics Workbench v.8.5.1 (evaluation version, CLC Bio), with a k-mer = 63 and a minimum contig length of 1000 bp. Subsequently, these contigs were aligned against the plastome of *Haematoxylum brasiletto* (NC\_026679) employing BLASTN (Altschul et al. 1990) with default search parameters. Then, the proper order of the aligned contigs was determined according to the reference. Joint areas and gaps between the contigs were checked and filled by mapping the high-quality paired-end reads.

For the remaining species, total gDNA was sequenced by genome skimming. The raw data was filtered using Trimmomatic v.0.32 (Bolger et al. 2014) with default settings. Then, paired-end reads of clean data were filtered and assembled into contigs using GetOrganelle (last accessed August 25, 2016) (Jin et al. 2019) with several GenBank published plastomes of Leguminosae (including *Libidibia coriaria*, NC\_026677; *Lotus japonicus*, NC\_002694; *Medicago truncatula*, NC\_003119; *Glycine max*, NC\_007942 and *Apios americana*, NC\_025909)as the seed sequences, calling Bowtie2 v.2.2.6 (Langmead and Salzberg 2012), BLAST+ (Camacho et al. 2009), and SPAdes v.3.10 (Bankevich et al. 2012). The *de novo* assembly graph was visualized and edited using Bandage Ubuntu dynamic v.0.8.0 (Wick et al. 2015), then a complete or nearly-complete circular plastome was generated for each sample.

To annotate the plastomes, we first constructed multiple genome alignments using MAUVE v.2.1.1 (Darling et al. 2004) in Geneious v.9.1.4 (Kearse et al. 2012), then aligned the plastomes to published plastomes with similar structure using MAFFT v.7.4.0 (Katoh and Standley 2013) with default parameters, coupled with manual adjustment. Protein-coding genes were double-checked by locating open reading frames using the “Find ORFs” function in Geneious. We used the online tRNAscan-SE service (Schattner et al. 2005) to improve the identification of tRNA genes. To detect the number of matched reads and the depth of coverage, raw reads were remapped to the assembled plastomes with Bowtie2.

*Sequence Alignment and Cleanup*

Because some plastomes of Leguminosae are hyper-variable in gene order and content (Magee et al. 2010; Gurdon and Maliga 2014; Martin et al. 2014), it is hard to rapidly extract and align these loci. Utilizing the Python library Biopython (Cock et al. 2009), we developed a new custom script “get\_annotated\_regions\_from\_gb.py” ([https://github.com/Kinggerm/PersonalUtilities/](https://github.com/Kinggerm/PersonalUtilities/;%20get_annotated_regions_from_gb.py)) to automatically extract all annotated regions and regions between annotations from legume plastomes. This is a command-line script that could be used to extract genes, CDSs, introns, and intergenic spacer regions from a list of annotated GenBank-format files. The output is separate FASTA files organized by regions (genes etc.) with a concomitant TAB-formatted file describing the lengths of each region of each sample. This script would be useful for extracting target loci from plastomes with high structural variations, and also applicable to plastomes maintaining a conserved configuration.

Each coding and noncoding locus was individually aligned using the L-INS-i method of MAFFT v.7.3.0 as implemented on the server platform. To minimize the use of loci with limited information or with relatively few species, loci consisting of fewer than four species were excluded, and aligned regions less than 22 bp were also removed. In addition, tRNAs were excluded from analyses because they are short (≤ 93 bp) and very conserved. We finally obtained 226 alignments including 81 coding and 145 noncoding loci.

*Datasets*

We first constructed three basic datasets: the PC (plastid coding regions; the concatenated 81 coding genes), PN (plastid noncoding regions; the concatenated 145 noncoding loci) and PCN (the concatenated PC and PN) datasets. To rapidly concatenate the alignments of separate loci and generate a concomitant configure file for use in downstream partition-based analysis, we also developed a new custom script “concatenate\_fasta.py”. This script should be useful in other phylogenomic studies.

Multiple strategies were then applied to reduce systematic error for the three basic datasets. The first strategy involved excluding ambiguously aligned regions using Gblocks v.0.91b (Castresana 2000; Talavera and Castresana 2007). Relaxed, default, and strict parameters (“Allowed Gap Positions” = “With All/Half/None”) were used to produce the “PC-GB-relaxed/default/strict”, “PN-GB-relaxed/default/strict” and “PCN-GB-relaxed/default/strict” datasets, respectively. The second strategy involved identifying and excluding loci with high levels of substitutional saturation. Linear regression of patristic distance and uncorrected *p* distances were used to determine the degree of saturation at 226 loci of 187 species by using TreSpEx v.1.1 (Struck 2014). Density plots (distributions) of the slope and *R*2 values were generated with R v.3.2.2 (R Development Core Team 2015). The rationale of TreSpEx is that the better the fit to linear regression, the less saturated the data are. The 47 loci (all noncoding loci, Supplementary Table S6) located on the left “hump” of the *R*2 distribution (determined as the value at which the distribution of the *R*2 values began to increase, or 0.54; Supplementary Fig. S1a) were pruned from the PCN dataset to form the PCN-*R2* dataset. The slope of the linear regression of each locus was also calculated, yielding two peaks with one valley between them (Supplementary Fig. S1b). Here, two thresholds for exclusion of loci were determined: a value of 0.180 representing the leftmost distribution of slopes, and a value of 0.423 representing the minimum between the two peaks in the graph (Supplementary Fig. S1b). The PCN-Slope1 dataset was produced by removing the 60 loci (including five coding regions: *accD*, *psbF*, *psbJ*, *psbL* and *psbM*) that fell within the leftmost distribution (blue area in Supplementary Fig. S1b), and the PCN-Slope2 dataset was generated by excluding the 124 loci (including 12 coding genes, see Supplementary Table S6) with slopes lower than 0.423 (corresponding to the areas in light blue and green in Supplementary Fig. S1b). The *R2* and slope values are listed in Supplementary Table S2. Most introns and intergenic regions were excluded in the strictest PCN-Slope2 dataset (Supplementary Table S6). Based on the cutoff value of *R2* and slope, the PC-*R*2/slope1/slope2 and PN-*R*2/slope1/slope2 datasets were generated. The third strategy involved excluding loci with low average bootstrap support, which has been shown to significantly reduce conflict (Salichos and Rokas 2013). Six sub-datasets (PC-BS70/80, PN-BS70/80 and PCN-BS70/80) were generated by only retaining loci with at least 70% and 80% ultrafast bootstrap (UFBoot) support, respectively, averaged across all nodes in the corresponding maximum likelihood tree. The criteria used to generate all corresponding datasets are presented in Table 1 and the loci included in each dataset are listed in Supplementary Table S6.

*Phylogenetic Analysis*

The best partitioning schemes (partitioned by locus) and substitution models were selected using PartitionFinder2 v.2.1.1 (Lanfear et al. 2017) with the all model, RAxML v.8.2.12 (Stamatakis 2014) and the rcluster algorithm (with the rcluster-percent set to 10; Lanfear et al. 2014), under the AICc criterion to compare the fit of the different models. The best partitioning scheme was then input into IQ-TREE v.1.6.5 (Nguyen et al. 2015; Chernomor et al. 2016) for model selection using the parameter TEST. All partitions shared the same set of branch lengths, but each could have a different substitution model and evolutionary rate via the -spp option. This was followed by 1000 independent likelihood searches from a random starting tree, each with a single ultrafast bootstrap (UFBoot2, Hoang et al. 2018) replicate and run in IQ-TREE. The trees for each locus were estimated in IQ-TREE followed by 1,000 likelihood tree searches, with the parameter TEST for model selection. We also inferred trees for each locus using RAxML v.8.1.2 including 1,000 bootstrap replicates and a search for the best scoring tree, employing the GTRGAMMA model. After these analyses, four datasets were excluded from subsequent analyses due the detection of an outlier tree topology (PN-GB-strict dataset, Fig. S7, S8), insufficient taxon sampling (PN-slope2 datasets, Table 1) or because the datasets (PCN-GB-strict, PCN-GB-slope2 datasets) included the former two problematic datasets.

*Quantification of Phylogenetic Signal for Alternative Tree Topologies*

At several key nodes in Leguminosae that resisted resolution among the datasets described above, we examined phylogenetic conflict following the methods of Smith et al. (2011), Shen et al. (2017) and Walker et al. (2018). Specifically, we tested phylogenetic signal within three sets of conflicting topologies across each of the 22 generated datasets: (1) alternative resolutions of the root of legumes, (2) alternative placements of *Griffonia*, and (3) alternative placements of *Pterogyne*. In this study, we recognized hard conflicts as those with conflicting bootstrap supports (UFBoot) ≥ 95% [IQ-TREE manual of Nguyen et al. (2015)]. The distribution of support for these three pairs of topologies was measured through comparisons of gene-wise log-likelihood support (GLS). For the first set of conflicting topologies, concerning the root of Leguminosae (Supplementary Fig. S2), we examined three hypotheses: (Detarioideae, (Cercidoideae, DDCP)), (T1, which was the constrained ML tree from the PCN-GB-relaxed dataset); ((Detarioideae, Cercidoideae), DDCP), (T2, which was the unconstrained ML tree from the PCN-GB-relaxed dataset); and (Cercidoideae, (Detarioideae, DDCP)), (T3, which was also the constrained ML tree versus T2). In order to quantify phylogenetic signal, we first calculated the site-wise log-likelihood scores (SLS) for T1, T2 and T3. Next, we calculated the difference in site-wise log-likelihood scores (ΔSLS) among T1, T2 and T3 for every site in a given dataset. By summing the ΔSLS scores of all sites for every gene in a given dataset, we then obtained the difference in gene-wise log-likelihood scores (ΔGLS) among T1, T2 and T3. By doing so, we were able to quantify the distribution of the phylogenetic signal for T1, T2 and T3 at the site and gene levels, as well as visualize the proportions of supporting sites or genes for T1, T2 and T3. We quantified the phylogenetic signal and visualized the proportion for the three alternative hypotheses for each of the abovementioned 22 datasets. Then, we used two measures to prune the three basic datasets (PC, PN, and PCN) in an effort to reduce the conflict at the root. First, because spurious inferences from loci with low information could lead to inaccurate or biased results, we removed the loci supporting alternative topologies in different datasets indicated by phylogenetic signal analyses. This involved removing 10 loci supporting different topologies from the PC dataset (=the PC-10-removed dataset), removing 26 loci supporting different topologies from the PN dataset (=the PN-26-removed dataset), and removing 36 loci from the PCN dataset (=the PCN-36-removed dataset). Second, because the ΔGLS is correlated with gene length in plastid genes, we calculated the average ΔSLS for each gene to avoid the effect of length in the PC, PN and PCN datasets and used standard deviation to identify outliers—that is, outlier loci were defined as those whose average ΔSLS values were greater than the upper bound or smaller than the lower bound of a Gaussian-like distribution: upper bound = min(max(*x*), μ + 3\*σ), lower bound = max(min(*x*), μ + 3\*σ), where max(*x*), min(*x*), μ, and σ are the maximum, minimum, average value, and standard deviation for a set of average ΔSLS values, respectively). This involved removing one outlier gene (*psbT* gene, Supplementary Fig. S3a) from the PC dataset to generate the PC-outlier-removed dataset, removing five outlier loci (5′-*trnM(CAU)*−5′-*trnV(UAC)* exon1, 5′-*psbJ*−3′-*psbL*, 5′-ccsA−3′-*trnL(UAG)*, 3′-*cemA*−5′-*petA*, 3′-*petG*−3′-*trnW(CCA)*, Supplementary Fig. S3b) from the PN dataset to generate the PN-outlier-removed dataset, and removing two outlier loci (*psbT* and *petL*, Supplementary Fig. S3c) from the PCN dataset to generate PCN-outlier-removed dataset. Phylogenetic trees were then reconstructed using IQ-TREE as described previously. Phylogenetic signal was also recalculated to compare the effect of the abovementioned two removals.

For the position of *Griffonia* (Supplementary Fig. S4a), we used three basic datasets to examine two alternative hypotheses (T1: *Griffonia* + *Bauhinia* *s*.*l*.; T2: *Griffonia* + *Bauhinia* *s*.*l*. I). For the position of *Pterogyne* (Supplementary Fig. S4b), similarly, we also examined two alternative hypotheses [T1: *Pterogyne* + the *Caesalpinia* clade; T2: *Pterogyne +* (*Dimorphandra* Group + the *Tachigali* clade + the *Peltophorum* clade + the Mimosoid clade) (abbreviated as *Pterogyne* + the DTPM clade)]. Site-wise log-likelihood scores were obtained for each topology using the -f G and the GTRGAMMA settings of RAxML.

*Test of Topological Concordance*

Topological concordance among phylogenetic trees for all 31 datasets (except the PN-slope2 dataset including limited taxa, Table 1) was estimated by all-to-all Robinson-Foulds distance using IQ-TREE and Principal Coordinates Analysis (PCoA) clustering in R v.3.3.3 (R Development Core Team 2015), which calculates the best reduced-spaced visualization of the distances between trees. In addition, topological concordance between phylogenetic trees was also estimated by Robinson-Foulds symmetric differences and the UPGMA clustering method using TreeSpace (Jombart et al. 2017).

Concordance among the 28 trees generated from the 28 different datasets was analyzed using PhyParts (Smith et al. 2015) and visualized using the ETE3 Python toolkit (Huerta-Cepas et al. 2016) as implemented in PhyParts\_PieCharts (<https://github.com/mossmatters/MJPythonNotebooks>; last accessed September 9, 2019).

To find the tree most concordant with the datasets analyzed, we iteratively used the tree from each of the 28 datasets as a reference tree for mapping the remaining trees. PhyParts requires rooted trees; thus all trees were rooted using a custom python script with *Quillaja saponaria* and *Polygala karensium* (members of Fabales) as outgroups. For these analyses, UFBoot cutoffs were implemented at both 95% and 80% (consistent with previous analyses); any tree node with support under the given bootstrap cutoff was regarded as uninformative for the reference tree node. We present results based on the UFBoot cutoff of 80% (Supplementary Fig. S5) as 95% was shown to be overly conservative; specifically we present the tree from the PCN dataset as this showed the highest levels of concordance with the other dataset trees.

Concordance among gene trees was analyzed by mapping the RAxML trees inferred from the 226 loci against the PCN dataset tree (Supplementary Fig. S6), also using PhyParts. We used the RAxML gene trees because support values in IQ-TREE-based trees are often inflated. The PCN tree was used because it was determined to be most concordant with the other datasets, as noted above. Because some gene trees were missing taxa, all 226 trees were rooted using the ‘pxrr’ function in Phyx (Brown et al. 2017), with the following ranked outgroups: *Quillaja saponaria*, *Suriana maritima*, *Cadellia pentastylis*, *Polygala sibirica*, *Polygala karensium*, *Eperua* sp., *Schnella trichosepala*, *Styphnolobium japonicum*, *Andira inermis*, *Arachis duranensis*, *Glycyrrhiza uralensis*, and *Faidherbia albida*. These rooted trees were then used in PhyParts analyses. Only bipartitions with > 70% bootstrap support were considered informative in the PhyParts analysis.

RESULTS

*Conflicting Phylogenetic Signals in Plastomes*

The multiple strategies we employed to reduce systematic error, while valuable in many cases, failed to reduce phylogenetic conflicts among datasets at some nodes, in particular for the three sets of conflicting relationships that we examined in detail. For these three cases of conflict, we quantified phylogenetic signal and visualized the proportions of genes in each dataset supporting the alternative topologies (Supplementary Fig 2, S2, S4 and Table S5).

Additionally, for the root of legumes, examination of the ΔGLS values among T1, T2 and T3 (Fig. 2a and Supplementary Fig. S2) revealed that T1 had a higher ΔGLS proportion of supporting genes (ranging from 42.73% to 56.40%; the ΔGLS values of datasets corresponding to these two proportions were 8.23 and 6.80, respectively) than those favoring either T2 or T3 in most PC and PC-derived datasets. T2 had a higher ΔGLS proportion of supporting genes than those favoring either T1 or T3 in most PN datasets (ranging from 37.31% to 49.22%; the ΔGLS values of datasets corresponding to these two proportions were 17.17 and 20.26, respectively), and T3 had a higher ΔGLS proportion of supporting genes than those favoring either T1 or T2 in some datasets of the PC and PCN trimmed by GBlocks (ranging from 37.73% to 51.78%; the ΔGLS values of datasets corresponding to these two proportions were 25.53 and 14.48, respectively). We further reanalyzed the distribution of ΔGLS to quantify the effect of removal of inconsistently supported loci and outlier loci in the three basic datasets (Fig. 2b). Removal of inconsistently supported loci and outlier loci did not change the topology among Detarioideae, Cercidoideae and DDCP (Supplementary Table S4). Removal of inconsistently supported loci only slightly increased the proportion of supporting genes (e.g., the ΔGLS proportion of genes supporting T1 increased from 44% to 47% in the PC dataset and from 32% to 33% in the PN dataset; Fig. 2b). Similarly, removal of outlier loci increased the ΔGLS proportion of genes supporting T1 just from 42% to 46% in the PC dataset; the ΔGLS proportion of genes supporting T2 increased from 38% to 40% in the PN dataset. Examination of the distribution of ΔGLS values also revealed that only a few genes displayed high ΔGLS values in the PC dataset (two genes with > |2lnL|) and PN dataset (two loci with > |2lnL|); moreover, these genes with high loci ΔGLS values favoring T1, T2, or T3 are randomly and unevenly distributed in the different regions of plastome. In other words, many genes are uninformative and their inclusion or exclusion does not greatly impact phylogenetic resolution. In addition, ΔGLS proportions favoring T1 and T2 are nearly equal in the PC and the PN datasets. Examination of the distribution of ΔSLS values in genes also revealed strong sites favoring T1, T2 or T3 unevenly and randomly distributed in the different loci of plastome.

For the positions of *Griffonia* and *Pterogyne*, our results showed that T1 had a higher proportion of supporting genes than those favoring T2 in the PC datasets, but that T2 had a higher proportion of supporting genes in the PN datasets (Supplementary Fig. S4). In the analysis of the position of *Griffonia*, 23 genes favoring T1 or T2 displayed high ΔGLS values with > |2lnL|, even > |10lnL| (e.g., *accD* and *ycf2*) in the PC dataset (Supplementary Table S5-29, Fig. S4a), and 59 loci favoring T1 or T2 displayed high ΔGLS values with > |2lnL|, even > |10lnL| (e.g., 3′-*petB* exon2−5′-*petD* exon1 and 5′-*trnL(UAA)* exon1−5′-*trnT(UGU)*) in the PN dataset (Supplementary Table S5-30, Fig. S4a). In the analysis of the position of *Pterogyne*, six genes favoring T1 or T2 displayed high ΔGLS values with > |2lnL| (e.g., *ndhC* and *rbcL*) in the PC dataset (Supplementary Table S5-32, Fig. S4b), and eight loci favouring T1 or T2 displayed high ΔGLS values with > |2lnL| (e.g., 3′-*psaJ*−5′-*rpl33* and 3′-*petB* exon1−5′-*petB* exon2) in the PN dataset (Supplementary Table S5-33, Fig. S4b).

For each pair of contrasting topologies, all loci showed ΔGLS values ranging from ~0 to 42.86, some loci displayed very high ΔGLS value in the analysis of the position of *Griffonia*, and more loci displayed high ΔGLS values than in the other two sets of contrasting topologies (Fig. 2, Supplementary Fig. S4). Similarly, these loci with strong signal favoring T1, T2 or T3 unevenly distributed in the different regions of plastome.

*Outlier Genes in the Plastome*

In phylogenetic signal analyses, some genes or fragments presented seemingly disproportionate ΔGLS values. In the PC dataset, *ycf1* and *ndhA* genes supporting T1 have higher ΔGLS (>2 lnL) but only moderate average ΔSLS. In the PN dataset, the 3′-*rps16* exon1−5′-*rps16* exon2 fragment supporting T2 and the 5′-*petL*−5′-*psbE* fragment supporting T1 had higher ΔGLS (>2 lnL) but only moderate average ΔSLS. Thus, the gene length is a major determinant of ΔGLS in phylogenetic signal analysis in the plastome, which is generally consistent with the findings of Walker et al. (2019) in terms of the significance of gene length. Our results also indicated the genes with disproportionate ΔGLS values should not be removed as outliers. Based on disproportionate average ΔSLS values, we removed *psbT* from the PC dataset; 5′-*trnM(CAU)*−5′-*trnV(UAC)* exon1, 5′-*psbJ*−3′-*psbL*, 5′-*ccsA*−3′-*trnL(UAG)*, 3′-*cemA*−5′-*petA*, 3′-*petG*−3′-*trnW(CCA)* from the PN dataset; and *psbT* and *petL* from the PCN dataset. These genes or fragments are short with moderate ΔGLS values. Removing these ‘outlier’ genes did not change the topology, only slightly changed bootstrap support, and did not significantly alter ΔGLS values.

*Concordance Analyses of Gene Trees*

Our analyses recovered the majority of clades with strong support (Fig. 1, Supplementary Fig. S5), but with some gene tree discordance (Supplementary Fig. S6). Most loci (~97%) were uninformative or in conflict at the root node of legume; 95% of loci were uninformative or in conflict at the position of *Griffonia*; and all loci were uninformative or in conflict at the position of *Pterogyne*. Concordance analyses of gene trees highlighted that most genes were uninformative for or conflicted with these particular nodal resolutions. In contrast, at the other ~68% of nodes, there were often more genes supporting than conflicting with the optimal species tree topology. However, the majority of plastid genes were uninformative for most nodes when considering support and thus unlikely to positively mislead studies. Insufficient information could lead to spurious tree inferences, thus producing noise/conflict. The study of Walker et al. (2019) suggested that stochastic error is a major source of conflict among plastid regions, especially among shorter or slower-evolving genes harboring insufficient signal to accurately infer relationships.

*New Insights into Deep Phylogenetic Relationships of Leguminosae*

With the exclusion of datasets that produced an outlier tree topology (PN-GB-strict dataset, Fig. S7, S8), had limited taxon sampling (PN-slope2 dataset, Table 1) and consisted of one of the first two (PCN-GB-strict, PCN-GB-slope2 datasets), the remaining 28 datasets produced largely congruent relationships of legumes, regardless of the type of sequences analyzed (coding vs. non-coding) and the different strategies to ameliorate systematic error by removing sites, loci, or outlier loci.

All subfamilies (excluding the monotypic Duparquetioideae) were recovered as monophyletic with strong support (UFBoot = 100%) in all analyses (Fig. 1, Supplementary Fig. S5 and Table S4). The clade (Duparquetioideae, (Dialioideae, (Caesalpinioideae, Papilionoideae))), abbreviated as DDCP, was strongly supported in all analyses, consistent with the results of LPWG (2017) based on *matK* and 81 plastid coding genes; this relationship was also recovered in the analyses of Koenen et al. (2020); except that Duparquetioideae was absent from their nuclear dataset. However, the relationships among DDCP, Cercidoideae and Detarioideae remained unresolved in our analyses. All three possible relationships among Cercidoideae, Detarioideae, and the DDCP clade were supported by different datasets in our analyses. The topology of (Cercidoideae, (Detarioideae, DDCP)) was strongly supported in the PC-GB-default dataset (UFBoot ≥ 94%) and the PCN-GB-default dataset (UFBoot ≥ 93%), consistent with previous studies based on a few plastid loci (e.g., Doyle et al. 2000; Bruneau et al. 2001; Kajita et al. 2001) as well as the plastome analyses of Koenen et al. (2020). The topology of (Detarioideae, (Cercidoideae, DDCP)) was supported by the PC (UFBoot ≥ 92%), PC-outlier-removed (UFBoot ≥ 95%), PC-BS70 (UFBoot ≥ 98%), PC-BS80 (UFBoot ≥ 97%), PC-10-removed (UFBoot ≥ 92%), PCN-BS70 (UFBoot ≥ 96%), and PCN-BS80 (UFBoot ≥ 94%) datasets; this relationship was also weakly supported in the study of Bruneau et al. (2008). The topology of ((Cercidoideae, Detarioideae), DDCP) was strongly supported by most PN-derived datasets with UFBoot ≥ 90% (Supplementary Table S4); the same topology was reconstructed based on 101 single-copy nuclear genes (Bootstrap Support = 61%; Cannon et al. 2015), while the nuclear analyses of Koenen et al. (2020) recovered Cercidoideae and Detarioideae as clade sister to DDCP. Our other analyses only weakly or moderately (UFBoot = 67−89%) supported one of the above three topologies. The difficulty in confidently resolving these deepest relationships of Leguminosae has been attributed to rapid diversification of these lineages (Lavin et al. 2005; Koenen et al. 2020) and ancient polyploidization (Cannon et al. 2015).

In contrast to these problematic deep relationships, our analyses significantly clarified relationships among major clades within Caesalpinioideae and Papilionoideae (Fig. 1, Supplementary Fig. S5 and Table S4). Within Caesalpinioideae, the two clades of the *Umtiza* grade, ((*Arcoa*, (*Acrocarpus*, *Ceratonia*)) and (*Umtiza*,(*Gleditsia*, *Gymnocladus*)), were subsequent sisters to remaining members of the subfamilyin all datasets except the PN-GB-default dataset. The *Umtiza* grade was weakly supported as a clade in Bruneau et al. (2008), whereas three unresolved clusters [namely *Arcoa*; (*Acrocarpus*, *Ceratonia*); and (*Umtiza*,(*Gleditsia*, *Gymnocladus*))] were supported in Manzanilla and Bruneau (2012). Our results were largely consistent with the analysis of 81 plastome genes in LPWG (2017), although *Arcoa* was not sampled in that study. After the *Umtiza* grade, a robustly supported *Cassia* clade was resolved as sister to the remaining Caesalpiniodieae. Within these remaining caesalpinioids, the position of *Pterogyne* varied among analyses. Seventeen datasets supported the *Caesalpinia* clade and *Pterogyne* as successive sisters to a clade formed by the *Dimorphandra* group, the *Tachigali* clade, the *Peltophorum* clade, and the Mimosoid clade (≥ 95% UFBoot) (Fig. 1 and Supplementary Table S4). Nevertheless,a clade composed of *Pterogyne* and the *Caesalpinia* clade was resolved as sister to a clade formed by the *Dimorphandra* group, the *Tachigali* clade, the *Peltophorum* clade, and the Mimosoid clade in three other datasets (UFBoot ≥ 97%) (Supplementary Table S4). *Pterogyne* and the *Caesalpinia* clade have been suggested to form a clade with the *Cassia* clade, within which *Pterogyne* is sister to the *Cassia* clade (Manzanilla and Bruneau 2012, Gagnon et al. 2016) or the *Caesalpinia* clade (Bruneau et al. 2008). The remaining Caesalpinioideae taxa were divided into two clades, one formed by the *Peltophorum* clade, the *Tachigali* clade and the *Dimorphandra* Group A taxa, and the other formed by the Mimosoid clade + the *Dimorphandra* Group B clade.

Within Papilionoideae, our study strongly supported the Swartzioid clade, the ADA clade, and the *Cladrastis* clade as successive sisters to the 50-kb inversion clade (Fig. 1), whereas previous studies recovered, with weak support, an ADA and Swartzioid clade as the first diverging lineage (Wojciechowski et al. 2004) or the ADA clade and the Swartzioid clade as successive sisters to remaining papilionoids (Cardoso et al. 2012; Cardoso et al. 2013). The phylogenetic positions of the *Andira* clade and the Lecointeoid clade had not been resolved in previous studies (e.g., Wojciechowski et al. 2004; Cardoso et al. 2012; Cardoso et al. 2013); these were strongly resolved as the first and second diverging lineages in the 50-kb inversion clade in our study. Phylogenetic relationships among the *Dermatophyllum*, *Vatairea*, Genistoid, Dalbergioid and Baphioid + Canavanine Accumulating (CA) clades had also not been well-resolved by previous studies (Cardoso et al. 2012; Cardoso et al. 2013). Within this group of taxa, our study mainly obtained three topologies: (*Vatairea*, (*Dermatophyllum*, ((Genistoid, Dalbergioid), (Baphioid, CA clade)))) was strongly supported by the PN-GB-relaxed dataset (Supplementary Fig. S12 and Table S4)); (*Vatairea*, (*Dermatophyllum*, (Genistoid, (Dalbergioid, (Baphioid, CA clade))))) was strongly supported by the PN-BS70 and PN-BS80 datasets; and ((*Vatairea*, *Dermatophyllum*), ((Genistoid, Dalbergioid), (Baphioid, CA clade))) was strongly supported by the PCN-GB-relaxed dataset (Supplementary Table S4).

Our study also clarified previously unresolved relationships within the other subfamilies. Previous studies had not resolved the relationships of the early-diverging Detarioideae (e.g., Bruneau et al. 2001; Bruneau et al. 2008; de la Estrella et al. 2017; de la Estrella et al. 2018). The six tribes recognized by de la Estrella et al. (2018) were each strongly supported and resolved into two strongly supported clades: ((Schotieae, (Barnebydendreae, Detarieae)) sister to ((Saraceae, (Afzelieae, Ameherstieae)). Within Amherstieae, *Amherstia* was resolved as sister to the *Brownea* clade in most datasets, or as sister to the *Berlinia* clade and some Ameherstieae taxa in the PN-BS70 and PCN-BS70 datasets (Supplementary Table S3). Within Cercidoideae, *Cercis* and *Adenolobus* were robustly supported as the first and second diverging lineages, which is consistent with the results from Bruneau et al. (2008) and Sinou et al. (2009). *Bauhinia* *s*.*l*. was resolved into two strongly supported clades, *Bauhinia* *s*.*l*. I (including *Bauhinia* and *Piliostigma*) and *Bauhinia* *s*.*l*. II (comprising *Barklya*, *Lysiphyllum*, *Tylosema* and *Schnella*) (Fig. 1). The placement of *Griffonia* was unresolved in past analyses, and in our analyses it was strongly supported as either sister to *Bauhinia* *s*.*l*. (by all PC datasets and some PCN datasets) or as sister to *Bauhinia* *s*.*l*. I (by all PN datasets and some PCN datasets; Fig. 1 and Supplementary Table S3). The latter relationship is consistent with that of Bruneau et al. (2008).

REFERENCES

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

Bankevich A., Nurk S., Antipov D., Gurevich A.A., Dvorkin M., Kulikov A.S., Lesin V.M., Nikolenko S.I., Pham S., Prjibelski A.D., Pyshkin A.V., Sirotkin A.V., Vyahhi N., Tesler G., Alekseyev M.A., Pevzner P.A. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19:455−477.

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

Brown J.W., Walker J.F., Smith S.A. 2017. Phyx: phylogenetic tools for unix. Bioinformatics 33:1886−1888.

Bruneau A., Forest F., Herendeen P.S., Klitgaard B.B., Lewis G.P. 2001. Phylogenetic relationships in the Caesalpinioideae (Leguminosae) as inferred from chloroplast *trnL* intron sequences. Syst. Bot. 26:487−514.

Bruneau A., Mercure M., Lewis G.P., Herendeen P.S. 2008. Phylogenetic patterns and diversification in the caesalpinioid legumes. Botany 86:697−718.

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

Cannon S.B., McKain M.R., Harkess A., Nelson M.N., Dash S., Deyholos M.K., Peng Y., Joyce B. Stewart Jr C.N., Rolf M., Kutchan T., Tan X., Chen C., Zhang Y., Carpenter E., Wong G.K., Doyle J.J., Leebens-Mack J. 2015. Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. Mol. Biol. Evol. 32:193−210.

Cardoso D., de Queiroz L.P., Pennington R.T., Lima H.C., Fonty E., Wojciechowski M.F., Lavin M. 2012. Revisiting the phylogeny of papilionoid legumes: new insights from comprehensively sampled early-branching lineages. Am. J. Bot. 99:1991−2013.

Cardoso D., Pennington R.T., de Queiroz L.P., Boatwright J.S., Van Wyk B.E., Wojciechowski M.F., Lavin M. 2013. Reconstructing the deep-branching relationships of the papilionoid legumes. S. Afr. J. Bot. 89:58−75.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17:540−552.

Chernomor O., von Haeseler A., Minh B.Q. 2016. Terrace sware data structure for phylogenomic inference from supermatrices. Syst. Biol. 65:997−1008.

Cock P.J.A., Antao T., Chang J.T., Chapman B.A., Cox C.J., Dalke A., Friedberg I., Hamelryck T., Kauff F., Wilczynski B., de Hoon M.J.L. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422−1423.

Darling A.C.E., Mau B., Blattner F.R., Perna N.T. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14:1394−1403.

de la Estrella M., Forest F., Wieringa J.J., Fougere-Danezan M., Bruneau A. 2017. Insights on the evolutionary origin of Detarioideae, a clade of ecologically dominant tropical African trees. New Phytol. 214:1722−1735.

de la Estrella M., Forest F., Klitgård B., Lewis G.P., Mackinder B.A., de Queiroz L.P., Wieringa J.J., Bruneau A. 2018. A new phylogeny-based tribal classification of subfamily Detarioideae, an early branching clade of florally diverse tropical arborescent legumes. Sci. Rep. 8:6884.

Doyle J.J., Doyle J.L. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bulletin. 19:11−15.

Doyle J.J., Chappill J.A., Bailey C.D., Kajita T. 2000. Towards a comprehensive phylogeny of legumes: evidence from *rbcL* sequences and non-molecular data. In: Herendeen P.S., Bruneau A., editors. Advances in Legume Systematics, Part 9. Richmond: Royal Botanic Gardens, Kew. pp. 1–20.

Gagnon E., Bruneau A., Hughes C.E., Queiroz L.P., Lewis G.P. 2016. A new generic system for the pantropical *Caesalpinia* group (Leguminosae). PhytoKeys 71:1−160.

Gurdon C., Maliga P. 2014. Two distinct plastid genome configurations and unprecedented intraspecies length variation in the *accD* coding region in *Medicago truncatula*. DNA Res. 21:417−427.

Hoang D.T., Chernomor O., von Haeseler A., Minh B.Q., Vinh L.S. 2018. UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35:518−522.

Huerta-Cepas J., Serra F., Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol. Biol. Evol. 33:1635–1638.

Jin J.-J., Yu W.-B., Yang J.-B., Song Y., dePamphilis C.W., Yi T.-S., Li D.-Z. 2019. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. bioRxiv 256479.

Jombart T., Kendall M., Almagro-Garcia J., Colijn C. 2017. Treespace: statistical exploration of landscapes of phylogenetic trees. Mol. Ecol. Resour. 17:1385−1392.

Kajita T., Ohashi H., Tateishi Y., Bailey C.D., Doyle J.J. 2001. *rbcL* and legume phylogeny, with particular reference to Phaseoleae, Millettieae, and allies. Syst. Bot. 26:515−536.

Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 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.

Koenen E.J.M., Ojeda D.I., Steeves R., Migliore J., Bakker F.T., Wieringa J.J., Kidner C., Hardy O.J., Pennington T.R., Bruneau A., Hughes C.E. 2020. Large-scale genomic sequence data resolve the deepest divergences in the legume phylogeny and support a near simultaneous evolutionary origin of all six subfamilies. New Phytol. 225:1355–1369.

Lanfear R., Calcott B., Kainer D., Mayer C., Stamatakis A. 2014. Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol. Biol. 14:82−95.

Lanfear R., Frandsen P.B., Wright A.M., Senfeld T., Calcott B. 2017. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol. Biol. Evol. 34:772−773.

Langmead B., Salzberg S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357−359.

Lavin M., Herendeen P.S., Wojciechowski M.F. 2005. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the Tertiary.Syst. Biol. 54:575−594.

LPWG [Legume Phylogeny Working Group]. 2013. Legume phylogeny and classification in the 21st century: progress, prospects and lessons for other species-rich clades. Taxon 62:217−248.

LPWG [Legume Phylogeny Working Group]. 2017. A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny. Taxon. 66:44−77.

Lewis G.P., Schrire B., Mackinder B., Lock M. 2005. Legumes of the World. Richmond: Royal Botanic Gardens, Kew.

Magee A.M., Aspinall S., Rice D.W., Cusack B.P., Semon M., Perry A.S., Stefanovic S., Milbourne D., Barth S., Palmer J.D., Gray J.C., Kavanagh T.A., Wolfe K.H. 2010. Localized hypermutation and associated gene losses in legume chloroplast genomes. Genome Res. 20:1700−1710.

Manzanilla V., Bruneau A. 2012. Phylogeny reconstruction in the Caesalpinieae grade (Leguminosae) based on duplicated copies of the sucrose synthase gene and plastid markers. Mol. Phylogen. Evol. 65:149−162.

Martin G.E., Rousseau-Gueutin M., Cordonnier S., Lima O., Michon-Coudouel S., Naquin D., de Carvalho J.F., Ainouche M., Salmon A., Ainouche A. 2014. The first complete chloroplast genome of the Genistoid legume *Lupinus luteus*: evidence for a novel major lineage-specific rearrangement and new insights regarding plastome evolution in the legume family. Ann. Bot*.* 113:1197−1210.

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.

Patel R.K., Jain M. 2012. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One 7:e30619.

R Development Core Team. 2015. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Salichos L., Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327−331.

Schattner P., Brooks A.N., Lowe T.M. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 33:W686−689.

Shen X.-X., Hittinger C.T., Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat. Ecol. Evol. 1:126−135.

Shved P., Mandrykin M., Mutilin V. 2012. Predicate Analysis with BLAST 2.7 (Competition Contribution). In: Flanagan C., König B., editors. TACAS 2012. LNCS. Heidelberg: Springer. pp. 525–527.

Sinou C., Forest F., Lewis G.P., Bruneau A. 2009. The genus *Bauhinia* s.l. (Leguminosae): a phylogeny based on the plastid *trnL-trnF* region. Botany 87:947−960.

Smith S.A., Wilson N.G., Goetz F.E., Feehery C., Andrade S.C., Rouse G.W., Giribet G., Dunn C.W. 2011. Resolving the evolutionary relationships of molluscs with phylogenomic tools. Nature 480: 364−367.

Smith S.A., Moore M.J., Brown J.W., Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. BMC Evol. Biol. 15:150−164.

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

Struck T.H. 2014. TreSpEx-detection of misleading signal in phylogenetic reconstructions based on tree information. Evol. Bioinform. 10:51−67.

Talavera G., Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56:564−577.

Walker J.F., Brown J.W., Smith S.A. 2018. Analyzing contentious relationships and outlier genes in phylogenomics. Syst. Biol. 67:916−924.

Walker J.F., Walker-Hale N., Vargas O.M., Larson D.A. Stull W.G. 2019. Characterizing gene tree conflict in plastome-inferred phylogenies. PeerJ. 7:e7747.

Wick R.R., Schultz M.B., Zobel J., Holt K.E. 2015. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics. 31:3350−3352.

Wojciechowski M.F., Lavin M., Sanderson M.J. 2004. A phylogeny of legumes (Leguminosae) based on analyses of the plastid *matK* gene resolves many well-supported subclades within the family. Am. J. Bot. 91:1846−1862.

Yang J.-B., Li D.-Z., Li H.-T. 2014. Highly effective sequencing whole chloroplast genomes of angiosperms by nine novel universal primer pairs. Mol. Ecol. Resour. 14:1024−1031.

Zhang T., Zeng C.-X., Yang J.-B., Li H.-T., Li D.-Z. 2016. Fifteen novel universal primer pairs for sequencing whole chloroplast genomes and a primer pair for nuclear ribosomal DNAs. J. Syst. Evol. 54:219−227.