CellPress
Partner Journal

Molecular Plant
Research Article

Nuclear phylotranscriptomics and phylogenomics support numerous polyploidization events and hypotheses for the evolution of rhizobial nitrogen-fixing symbiosis in Fabaceae

Yiyong Zhao[1,2], Rong Zhang[3], Kai-Wen Jiang[4,5], Ji Qi[1], Yi Hu[2], Jing Guo[1], Renbin Zhu[6], Taikui Zhang[1], Ashley N. Egan[7], Ting-Shuang Yi[3,*], Chien-Hsun Huang[1,*] and Hong Ma[2,*]

[1]State Key Laboratory of Genetic Engineering and Collaborative Innovation Center of Genetics and Development, Ministry of Education Key Laboratory of Biodiversity and Ecological Engineering, Institute of Plant Biology, Center of Evolutionary Biology, School of Life Sciences, Fudan University, 2005 Songhu Road, Shanghai 200433, China

[2]Department of Biology, The Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA

[3]Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Lanhei Road, Kunming 650201, China

[4]Key Laboratory of Biodiversity Conservation in Southwest China, State Forestry Administration, Southwest Forestry University, Kunming 650224, PR China

[5]Ningbo Botanical Garden Herbarium, Ningbo 315201, PR China

[6]Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Menglun, Yunnan 666303, PR China

[7]Department of Biology, Utah Valley University, Orem, UT 84058, USA

***Correspondence: Ting-Shuang Yi (tingshuangyi@mail.kib.ac.cn), Chien-Hsun Huang (huang_ch@fudan.edu.cn), Hong Ma (hxm16@psu.edu)**

https://doi.org/10.1016/j.molp.2021.02.006

ABSTRACT

Fabaceae are the third largest angiosperm family, with 765 genera and ~19 500 species. They are important both economically and ecologically, and global Fabaceae crops are intensively studied in part for their nitrogen-fixing ability. However, resolution of the intrasubfamilial Fabaceae phylogeny and divergence times has remained elusive, precluding a reconstruction of the evolutionary history of symbiotic nitrogen fixation in Fabaceae. Here, we report a highly resolved phylogeny using >1500 nuclear genes from newly sequenced transcriptomes and genomes of 391 species, along with other datasets, for a total of 463 legumes spanning all 6 subfamilies and 333 of 765 genera. The subfamilies are maximally supported as monophyletic. The clade comprising subfamilies Cercidoideae and Detarioideae is sister to the remaining legumes, and Duparquetioideae and Dialioideae are successive sisters to the clade of Papilionoideae and Caesalpinioideae. Molecular clock estimation revealed an early radiation of subfamilies near the K/Pg boundary, marked by mass extinction, and subsequent divergence of most tribe-level clades within ~15 million years. Phylogenomic analyses of thousands of gene families support 28 proposed putative whole-genome duplication/whole-genome triplication events across Fabaceae, including those at the ancestors of Fabaceae and five of the subfamilies, and further analyses supported the Fabaceae ancestral polyploidy. The evolution of rhizobial nitrogen-fixing nodulation in Fabaceae was probed by ancestral character reconstruction and phylogenetic analyses of related gene families and the results support the hypotheses of one or two switch(es) to rhizobial nodulation followed by multiple losses. Collectively, these results provide a foundation for further morphological and functional evolutionary analyses across Fabaceae.

Key words: Fabaceae, Leguminosae, nuclear phylogeny, divergence times, whole-genome duplication, rhizobial nodulation

Zhao Y., Zhang R., Jiang K.-W., Qi J., Hu Y., Guo J., Zhu R., Zhang T., Egan A.N., Yi T.-S., Huang C.-H., and Ma H. (2021). Nuclear phylotranscriptomics and phylogenomics support numerous polyploidization events and hypotheses for the evolution of rhizobial nitrogen-fixing symbiosis in Fabaceae. Mol. Plant. **14**, 748–773.

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.

INTRODUCTION

The legume family (Fabaceae/Leguminosae) is one of the most prominent angiosperm families across global ecosystems, with 765 genera and ~19 500 species (LPWG, 2017; Legume Phylogeny Working Group [LPWG]); it is the third largest angiosperm family after Asteraceae and Orchidaceae. Fabaceae are widely distributed, contribute to both tropical and temperate ecosystems, and include many economically important species (Lewis et al., 2005). The family exhibits a spectacular range of morphological and habit diversity, from giant rainforest trees and woody lianas to desert shrubs, ephemeral herbs, and herbaceous climbers. Legumes are used as high-nutrient grains, oils, vegetables, and forage, high-quality lumber, and medicinal herbs; they include soybeans, common beans, lima beans, mung beans, alfalfa, fava beans, peanuts, peas, chickpeas, lentils, and honey locust. Also, soybean (*Glycine max*), barrelclover (*Medicago truncatula*), and *Lotus japonicus* are important model plants for studying the molecular basis of legume traits (Barker et al., 1990; Handberg and Stougaard, 1992; Zhang et al., 2015). Most legumes can fix nitrogen with rhizobial bacteria in symbiotic root nodules, contributing to the great ecological success of the family. Because nitrogen is a major plant nutrient, the symbiosis of legumes with rhizobia in root nodules impacts the global nitrogen cycle and crop production (Zahran, 1999; Garg and Geetanjali, 2009; Dos Santos et al., 2012). In addition to legumes, members of eight other families from one of four orders in a large rosid clade (referred to as the "nitrogen-fixing clade") also contain nodule-producing species (Soltis et al., 1995; Sprent, 2009; Doyle, 2011, 2016; van Velzen et al., 2019).

Fabaceae were traditionally divided into three subfamilies, mainly according to their flower structures (Lewis et al., 2005). Legume phylogenies have been inferred using several chloroplast markers (Wojciechowski et al., 2004; Lavin et al., 2005; Bruneau et al., 2008; Cardoso et al., 2012, 2013). Recent analyses using chloroplast sequences (*matK* [3696 legumes, LPWG, 2017], the plastome [179 legumes in 160 genera, Zhang et al., 2020b]) or nuclear genes (7631 nuclear genes from 76 taxa, including 42 legumes with transcriptomes and genomes representing 5 subfamilies, Koenen et al., 2020a) support a new classification system that contains 6 subfamilies (Caesalpinioideae, Cercidoideae, Detarioideae, Dialioideae, Duparquetioideae, and Papilionoideae) with morphological support (LPWG, 2017; Koenen et al., 2020a; Zhang et al., 2020b). Caesalpinioideae (148 genera, ca. 4400 species; e.g., *Gleditsia* [honey locust]) are trees, shrubs, lianas, and perennials. Caesalpinioideae include a large clade, the mimosoids, which was previously the subfamily Mimosoideae (ca. 41 genera, 3300+ species; e.g., *Acacia* and *Mimosa*). Cercidoideae (12 genera, ca. 335 species; e.g., *Bauhinia* and *Cercis*) are often trees, shrubs, or lianas. Detarioideae (81 genera, ca. 760 species; e.g., *Afzelia* and *Macrolobium*) are usually trees. Dialioideae (17 genera, ca. 85 species; e.g., *Dialium*) are often trees or shrubs. Duparquetioideae (1 genus, 1 species: *Duparquetia orchidacea*) are scrambling lianas native to West Africa with unusual floral characteristics (Prenner and Klitgaard, 2008). Papilionoideae (503 genera, ca. 14 000 species; e.g., *Dalbergia* and *Phaseolus*) include many crop species (e.g., those mentioned above), as well as trees, shrubs, lianas, herbs, and twining vines. Among legumes, the ability to form symbiotic rhizobial nodules with nitrogen-fixing bacteria is found only in members of the two largest subfamilies, Papilionoideae and Caesalpinioideae (Doyle, 2016). Phylogenetic relationships among the subfamilies Cercidoideae, Detarioideae, and Duparquetioideae, and between them and Dialioideae are still not well resolved (LPWG, 2017; Koenen et al., 2020a; Zhang et al., 2020b) (Supplemental Figure 1A, 1B, and 1G).

Three of the subfamilies, Detarioideae, Caesalpinioideae, and Papilionoideae, are further subdivided into tribes, each with one or more genera, but molecular phylogenetic analyses suggest that some tribes and genera are not monophyletic (e.g., Egan et al., 2016; LPWG, 2017). Moreover, several (non-ranked) clades containing multiple genera (sometimes from more than one tribe) have been named with support from molecular phylogeny (LPWG, 2013; LPWG, 2017). However, the relationships among some tribes, clades, or genera remain unclear (Bruneau et al., 2008; LPWG, 2017). For example, in Detarioideae, the relationships of the tribes Detarieae, Barnebydendreae, and Schotieae are unresolved (de la Estrella et al., 2018). In addition, in the largest tribe Detarieae with 21 genera, the relationships among 8 genera are uncertain, as is the placement of *Goniorrhachis* (LPWG, 2017; de la Estrella et al., 2018) (Supplemental Figure 1C and 1D). In Caesalpinioideae, the mimosoid clade consists of three paraphyletic tribes, Acacieae, Ingeae, and Mimoseae, and includes two non-monophyletic genera, *Mimosa* and *Senegalia* (Luckow et al., 2003; Bouchenak-Khelladi et al., 2010; Kyalangalilwa et al., 2013). Among other taxa of Caesalpinioideae, several genera/lineages have unresolved relationships (Supplemental Figure 1F) (Bruneau et al., 2001; Bruneau et al., 2008; LPWG, 2017).

Many phylogenetic uncertainties remain in the largest subfamily, Papilionoideae. For example, the relationships of the species-poor sister lineages (SPSLs) to other members of Papilionoideae are uncertain, with either the tribe Swartzieae or the ADA clade (with the three tribes Amburaneae, Dipterygeae, and Angylocalyceae) as the sister of the remaining papilionoids (Supplemental Figure 1E and 1I) (Cardoso et al., 2013; LPWG, 2017; Zhang et al., 2020b). Also, relationships among many subclades are not resolved within the largest clade, the "50-kb inversion clade," (Supplemental Figure 1E and 1I), which includes the *Andira* and vataireoid clades, the tribe Exostyleae, the genistoids s.l., the dalbergioids s.l., and the NPAAA (non-protein amino acid-accumulating) clade. The NPAAA clade includes species whose seeds accumulate the non-proteinogenic amino acid canavanine, a deterrent against herbivory (Bell, 1981). It is also called the CA clade for canavanine accumulating [Wojciechowski et al., 2004]), and it contains most of the agriculturally cultivated legumes. In addition, relationships among some subclades within the NPAAA clade are also unclear (Supplemental Figure 1I) (LPWG, 2017). Moreover, several tribes or clades in Papilionoideae are not monophyletic, such as Millettieae and Phaseoleae (Egan et al., 2016) and the vataireoids (Supplemental Figure 1I) (Stefanović et al., 2009; Cardoso et al., 2013; LPWG, 2013).

A robust and well-resolved Fabaceae phylogeny is fundamental to understanding the aspects of Fabaceae evolution and diversification, including the relationships between crops and their wild

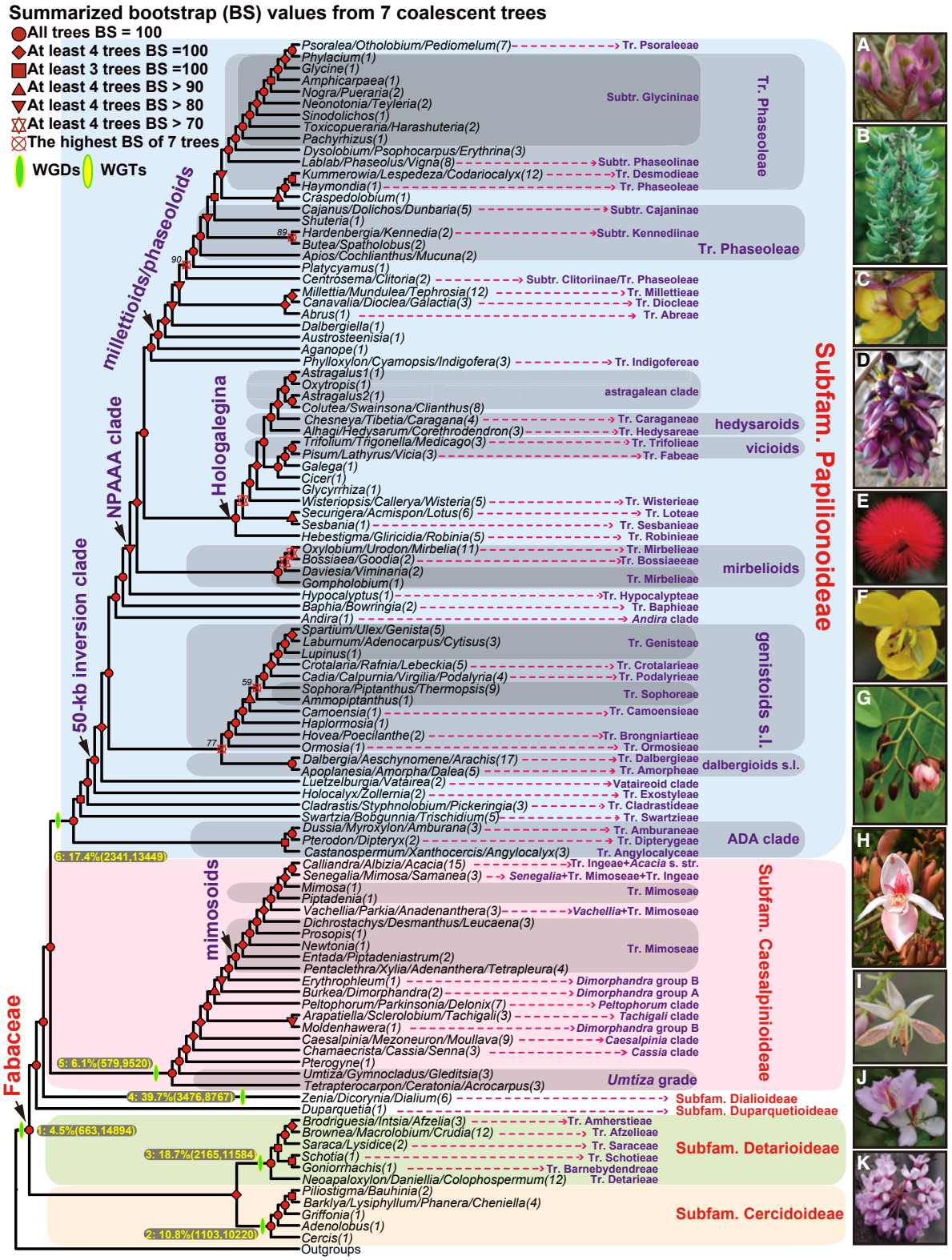

Figure 1. An overview of the Fabaceae phylogeny.

This is a summary tree from seven coalescent trees, simplified to present the major lineages with tips collapsed as noted in the tip names. A complete summary tree from the seven coalescent trees is provided in Figures 2, 3, 4, and 5 and Supplemental Figure 3. Numbers after the tips show the total number of genera sampled for each collapsed tip. Colored labels at the nodes represent the summarized status of bootstrap values in our seven coalescent trees: red circles for 100% BS in all seven trees; red diamonds for 100% BS in at least four trees; red squares for 100% BS in at least

(legend continued on next page)

relatives, the evolutionary history of key functional traits such as rhizobial nodulation, and the analysis of genes that underlie such traits. Furthermore, a reliable species tree is a crucial reference for the protection of wild germplasms, the preservation of biodiversity, and the planning of nature reserves. Recent advances in sequencing have greatly facilitated the use of low-copy nuclear genes for understanding plant phylogenies (Wen et al., 2015). Nuclear genes are inherited biparentally and usually show higher substitution rates than plastid genes (Birky, 2001; Springer et al., 2001; Davis et al., 2014; Lu et al., 2018). In addition, nuclear genes provide additional information that is useful for resolving incongruences between phylogenies that may arise from horizontal gene transfer, hybridization, rapid radiation, and incomplete lineage sorting (Zeng et al., 2014, 2017; Zhao et al., 2016; Li et al., 2017). Analyses using nuclear genes, especially phylotranscriptomics, have been successful in resolving relationships among major angiosperm lineages (Zhang et al., 2012; Wickett et al., 2014; Zeng et al., 2014; LPWG, 2017; Zeng et al., 2017; Yang et al., 2018; Zhang et al., 2020a), large families (Huang et al., 2016a, 2016b; Xiang et al., 2017; Yang et al., 2018; Mandel et al., 2019), and ferns (Shen et al., 2017; Qi et al., 2018). Transcriptomic/genomic datasets have also been generated for 20 (Cannon et al., 2015), 30 (Vatanparast et al., 2018), or 42 legume species (Koenen et al., 2020a) for analyzing phylogenetic relationships among subfamilies and some tribes.

Nuclear genes are also useful to study whole-genome duplication (WGD), or polyploidization, which is widespread in angiosperm history and is thought to contribute to functional innovation and morphological diversification (Freeling and Thomas, 2006; Jiao et al., 2011, 2014; Huang et al., 2016b; Ren et al., 2018; Yang et al., 2018; Leebens-Mack et al., 2019; Guo et al., 2020). Specifically, WGDs in Fabaceae are supported by syntenic regions in legume genomes (Cannon et al., 2006; Schmutz et al., 2010; Hane et al., 2017; Wang et al., 2017; Stai et al., 2019; Zhuang et al., 2019) and by clusters of gene duplications detected by phylotranscriptomic analyses, giving rise to hypotheses regarding autopolyploidy and allopolyploidy (Leebens-Mack et al., 2019; Stai et al., 2019; Koenen et al., 2020b). It is possible that the evolutionary and ecological success of Fabaceae has benefited from additional WGDs not yet uncovered in previous analyses. Further investigation of WGD in Fabaceae with phylogenetic placement and examination of different hypotheses could be facilitated by large-scale sequence datasets from greater numbers of legumes than have been examined previously.

In this study, we obtained transcriptomes and genomes from 463 Fabaceae species representing 333 genera, including newly sequenced transcriptomes and genomes for 391 legumes, and we identified 1559 low-copy nuclear genes for phylogenetic analyses using coalescent and supermatrix approaches. The resulting Fabaceae phylogeny provides consistently strong support for relationships among subfamilies, tribes/clades, and most of the genera sampled here. Moreover, molecular clock estimation suggests that the six subfamilies diverged within a few million years near the K/Pg boundary. Our phylogenomic analyses detected numerous clusters of gene duplications (GDs) in the legume phylogeny and provided support and phylogenetic positions for 28 WGD/whole-genome triplication (WGT) events, including those in the ancestors of the family and most subfamilies. We also present analyses and discussion of recently proposed hypotheses for polyploidy in early Fabaceae history (Koenen et al., 2020b). Furthermore, we investigated rhizobial nitrogen-fixing nodulation by ancestral state reconstruction and evolutionary analyses of related gene families. The results support the hypotheses of one or two switch(es) to rhizobial nodulation followed by multiple losses of nitrogen-fixing nodulation in legumes.

RESULTS AND DISCUSSION

Taxon sampling and selection of gene markers for phylogenetic analyses

We sampled 463 legumes (Supplemental Table 1, with authorship information) from all 6 subfamilies (307 Papilionoideae, 100 Caesalpinioideae, 35 Detarioideae, 13 Cercidoideae, 7 Dialioideae, and 1 Duparquetioideae), with 377 transcriptomes and 14 genomes newly generated here (Supplemental Table 2). We used a multi-step procedure to screen thousands of low-copy nuclear gene sets (referred to as orthogroups, or OGs) to obtain 7 OG sets of 1559, 1083, 982, 593, 484, 384, and 131 OGs (Supplemental Table 3) (see Methods, Supplemental Figure 2 and Supplemental text for details). For the 1559 OGs, the corresponding gene IDs in model plants are shown in Supplemental Table 4, and gene ontology enrichment analysis shows that these genes are highly conserved and have (or are predicted to have) mainly house-keeping functions such as DNA repair, RNA splicing, protein binding, nucleotide binding, rRNA binding, and methyl transfer (Supplemental Table 4).

Strongly supported relationships among Fabaceae subfamilies

We used the ASTRAL coalescent method to reconstruct legume phylogenies with the seven OG sets, resulting in trees that are largely consistent with each other (Figure 1, for details, see Figures 2, 3, 4, and 5 and Supplemental Figures 3 and4). Also, the maximum likelihood (ML) tree inferred from the concatenated dataset of the 131 OGs from the seventh gene set is largely consistent with the 7 coalescent trees (Supplemental Figure 5). The legume phylogeny is highly resolved, with strong support for the vast majority of relationships among taxa: 85.6% of the nodes have 100%

three trees; red triangles for >90% BS in at least four trees; inverted red triangles for >80% BS in at least four trees; red hexagons for > 70% BS in at least four trees; and red circles with an "X" for other types. The highest BS percent value among the seven coalescent trees is shown near the node. Photographs with a number on the right represent different subfamilies: **(A)** *Oxytropis bicolor*, **(B)** *Strongylodon macrobotrys*, **(C)** *Cajanus cajan*, and **(D)** *Mucuna sempervirens* for Papilionoideae; **(E)** *Calliandra haematocephala* and **(F)** *Chamaecrista mimosoides* for Caesalpinioideae; **(G)** *Zenia insignis* for Dialioideae; **(H)** *Duparquetia orchidacea* for Duparquetioideae; **(I)** *Tamarindus indica* for Detarioideae; and **(J)** *Phanera variegata* and **(K)** *Cercis glabra* for Cercidoideae. The green and yellow ellipses indicate large-scale gene duplications and triplications from phylogenomic/phylotranscriptomic analyses, with percentages, numbers of duplicated genes, and the total number of gene families (or clades related to the node) in gene trees shown adjacent to each node.

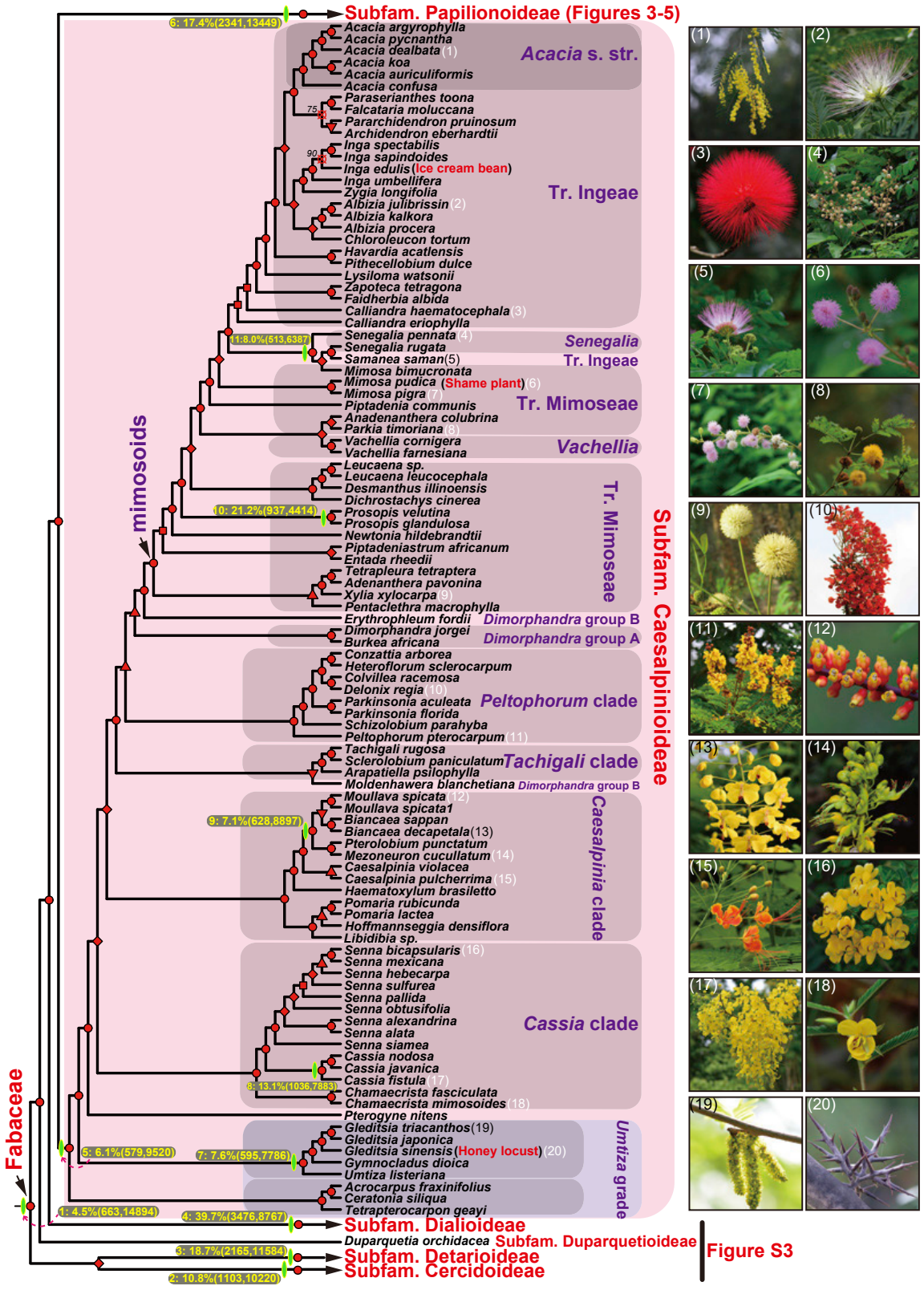

Figure 2. A summarized phylogeny of Caesalpinioideae from seven coalescent trees at the species level.
Definitions of red labels and ellipses on each node are the same as in Figure 1. The numbered plant photographs at the right of the figure show the species with corresponding numbers after their names.

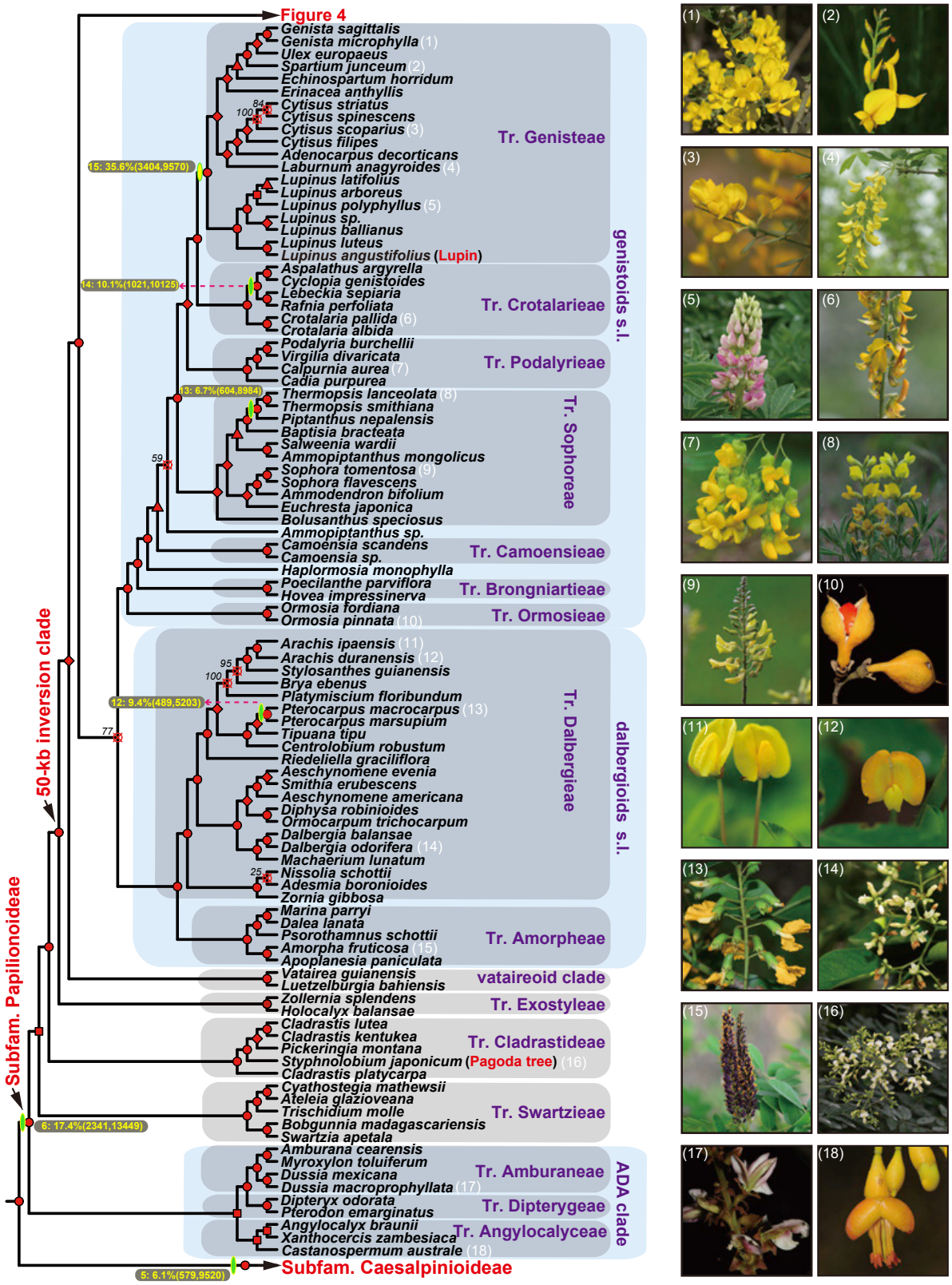

Figure 3. A portion of the summarized phylogeny of Papilionoideae with species-poor sister lineages of other Papilionoideae from seven coalescent trees at the species level.

Definitions of red labels and ellipses on each node are the same as in Figure 1. The numbered plant photographs at the right of the figure show the species with corresponding numbers after their names.

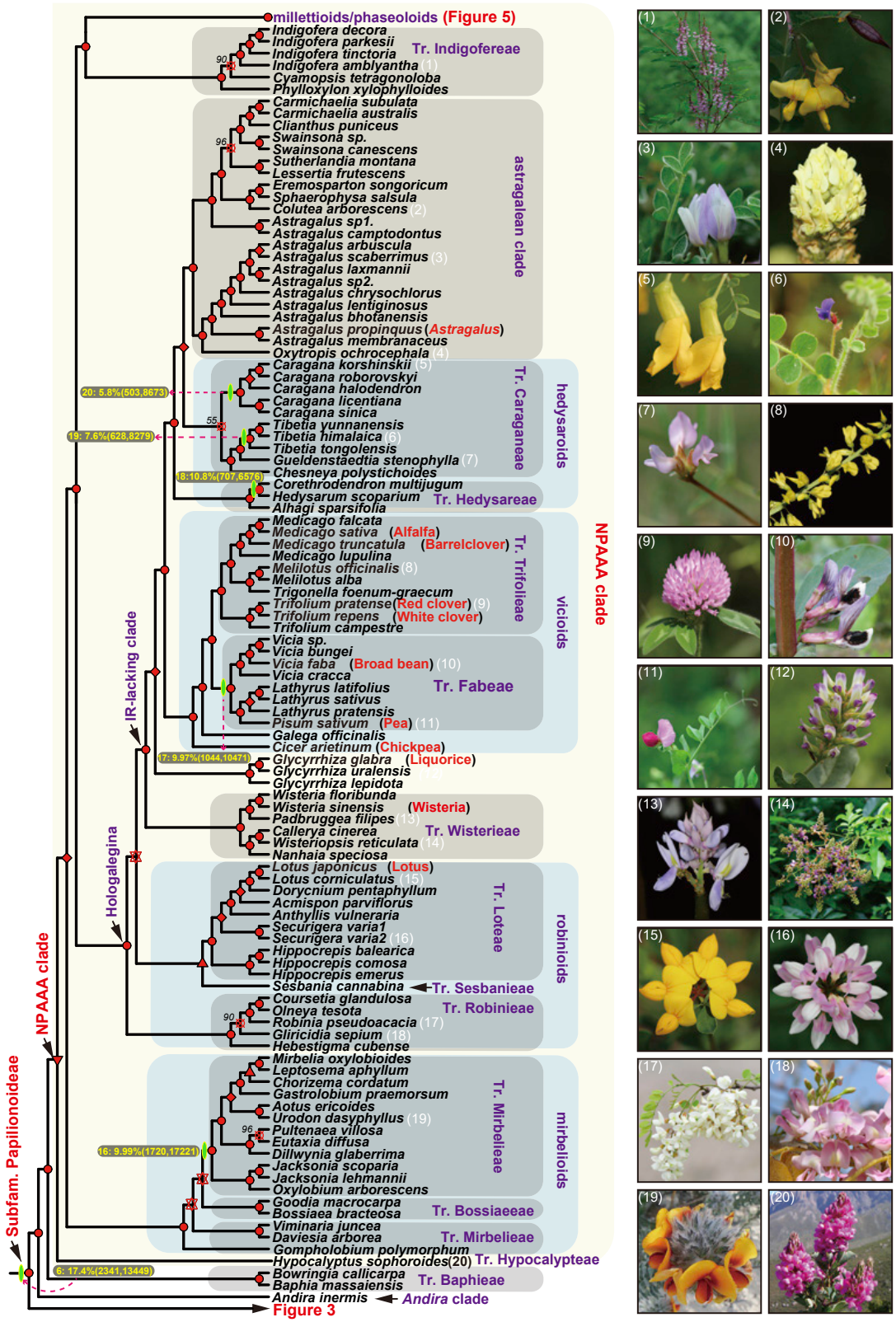

Figure 4. A summarized phylogeny of the NPAAA clade of Papilionoideae from seven coalescent phylogenetic trees at the species level.

Definitions for red labels and ellipses on each node are the same as in Figure 1. The numbered plant photographs at the right of the figure show the species with corresponding numbers after their names.

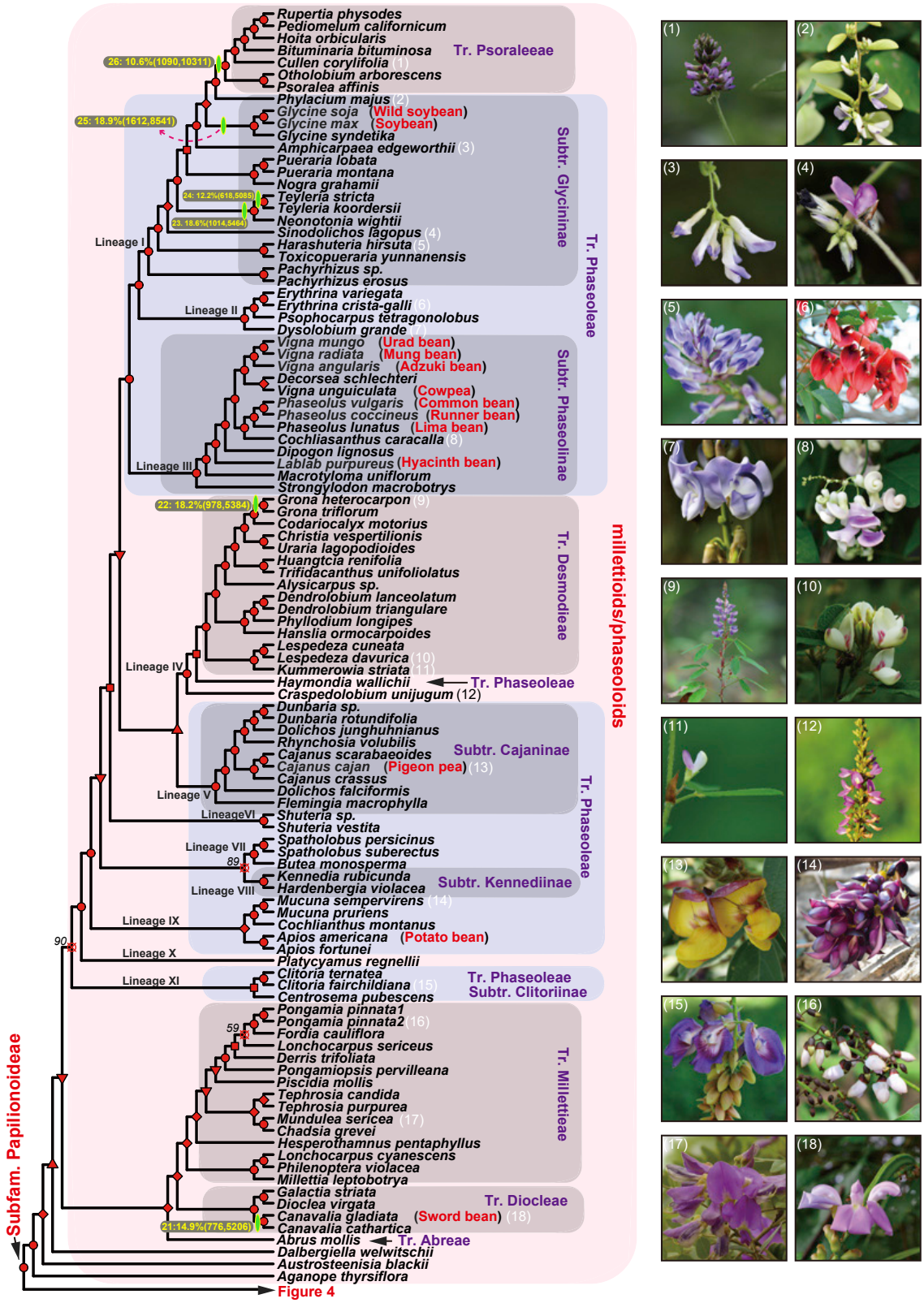

Figure 5. A summarized phylogeny of the millettioids/phaseoloids (Papilionoideae) from seven coalescent trees at the species level.
Definitions of red labels and ellipses on each node are the same as in Figure 1. The numbered plant photographs at the right of the figure show the species with corresponding numbers after their names.

bootstrap support (BS) values in all seven coalescent trees, and 91.3% of the nodes have at least 90% BS values (Supplemental Figure 4). Specifically, the monophyly of Fabaceae and five subfamilies (except the monotypic Duparquetioideae) is fully supported in coalescent analyses using the largest 3 OG sets with 1559, 1083, and 982 genes. These were obtained using only taxon coverage and gene length filters, but no additional steps such as the removal of putative paralogs (Supplemental Figure 2). The monophyly of the five subfamilies is consistent with previous and recent phylogenies based on chloroplast and nuclear genes (LPWG, 2017; Koenen et al., 2020a; Zhang et al., 2020b). Further phylogenetic analyses were performed using four other OG sets that were obtained using more filters (see Supplemental text, Figure 1 and Supplemental Figure 2).

The relationships among six Fabaceae subfamilies were resolved robustly with nearly 100% BS in all seven trees. Here, Cercidoideae and Detarioideae together form a clade that is sister to the other subfamilies, but in previous topologies, their relationships relative to other subfamilies were unresolved or uncertain (LPWG, 2017; Koenen et al., 2020a; Zhang et al., 2020b) (Supplemental Figure 1). Among the four other subfamilies, Duparquetioideae is sister to the other three subfamilies, and Dialioideae are sister to a clade that contains the two largest subfamilies, Caesalpinioideae and Papilionoideae. These relationships have strong support in all seven trees (summarized in Figure 1) and are consistent with recent studies (LPWG, 2017; Koenen et al., 2020a; Zhang et al., 2020b) (Supplemental Figure 1).

The 3 coalescent trees inferred from 1559, 1083, and 982 OGs (Supplemental Figure 2) also provide strong support for the monophyly of groups within subfamilies. Among the 59 tribes or rank-free clades here, 52 are represented by 2 or more taxa; 45 of the latter are monophyletic, whereas 7 are not (the tribes Ingeae, Mimoseae, Sophoreae, Phaseoleae, and Mirbelieae; *Senegalia*; and the *Dimorphandra* group B; Figures 2, 3, 4, and 5, and Supplemental Figure 3). Furthermore, the final summarized Fabaceae phylogeny (Figure 1) with results from all seven coalescent trees also supports the non-monophyly of the same seven tribes or rank-free clades mentioned above. These consistent relationships indicate the robustness of our phylogenetic results (Figures 1, 2, 3, 4, and 5, and Supplemental Figures 3 and 4).

Highly supported relationships within the subfamilies Cercidoideae, Detarioideae, and Dialioideae

Thirteen species representing 9 genera of Cercidoideae are included here (Supplemental Figure 3). The placement of *Griffonia* was previously uncertain (LPWG, 2017). Here, the use of hundreds of nuclear genes from 9 genera consistently placed *Griffonia* as sister to a clade that contained *Bauhinia* and other genera (Figure 1 and Supplemental Figure 3), similar to the relationships reported in a recent study that focused on this subfamily using chloroplast genes from taxa in seven genera (Wang et al., 2018).

In Detarioideae, 35 species from 14 genera are sampled here, representing all 6 previously defined tribes (de la Estrella et al., 2018). They are resolved into 2 clades, (1) Detarieae and (2) the Amherstieae clade that contains the tribes Amherstieae, Afzelieae, Saraceae, Barnebydendreae, and Schotieae (Figure 1 and Supplemental Figure 3), with 100% BS for all relationships based on the 3 largest sets of OGs (Supplemental Figure 4). However, phylogenies using smaller sets of OGs are inconsistent regarding relationships among these 5 tribes and have BS values ranging from 47 to 71, suggesting possible complex histories of nuclear genes (see Supplemental Figures 3 and 4 and Supplemental text for more information and a comparison with earlier topologies).

Dialioideae are represented here by seven species from six genera. Our phylogeny places the genera *Poeppigia* and *Baudouinia* as successive sisters to a clade of the other four genera (100% BS, Supplemental Figure 3), which form two smaller clades, *Labichea* + *Zenia* (at least four trees with BS = 100%) and *Dialium* + *Dicorynia* (all 100% BS), nearly consistent with relationships determined using a combined molecular and morphological dataset (Zimmerman et al., 2017). Although the relationships among *Baudouinia*, *Labichea*, and others were not resolved previously (Herendeen et al., 2003; Bruneau et al., 2008; LPWG, 2017), a recent analysis using whole plastomes also supported *Poeppigia* and *Baudouinia* as successive sisters of a clade with the other genera, but *Zenia* was placed in a subclade with *Dialium* and *Dicorynia* rather than being close to *Labichea* (Zhang et al., 2020b).

Highly supported relationships in Caesalpinioideae

The sampling of Caesalpinioideae here includes 100 species from 66 genera, representing major lineages at the tribe or equivalent level (Figure 2). Our phylogeny supports (100% BS) the placement of the small *Ceratonia* and *Gleditsia* clades, members of the "*Umtiza* grade," as successive sisters to the maximally supported clade of all remaining sampled Caesalpinioideae, consistent with relationships based on the analysis of whole plastomes (Zhang et al., 2020b). Here, *Tetrapterocarpon* is sister to *Acrocarpus* + *Ceratonia*, whereas *Umtiza* is sister to *Gleditsia* + *Gymnocladus*.

Among the subgroups in the large clade of remaining Caesalpinioideae, *Pterogyne*, the *Cassia* clade, and the *Caesalpinia* clade are successive sisters to the large clade of other Caesalpinioideae members (Figure 2), and both the *Cassia* and *Caesalpinia* clades are monophyletic (100% BS). The monophyly of the *Cassia* and *Caesalpinia* clades was demonstrated previously, but *Pterogyne* was placed as sister to either the *Cassia* clade (Manzanilla and Bruneau, 2012), the *Caesalpinia* clade (Bruneau et al., 2008; LPWG, 2017), or the other Caesalpinioideae (Bruneau et al., 2008; Manzanilla and Bruneau, 2012; Zhang et al., 2020b). Within the *Cassia* clade, *Chamaecrista* is sister to the clade of *Cassia* and *Senna*; in the *Caesalpinia* clade, the nine genera sampled here are resolved into two maximally supported clades with 3 and 6 genera, respectively (Figure 2). Four well-resolved clades, including members of the previously defined *Dimorphandra* groups A and B, are resolved as successive sisters to the large clade of mimosoids (Figure 2; see the Supplemental text for additional information). The topology here is consistent in all 7 trees (Supplemental Figure 4), and most nodes receive 80%–100% BS.

In mimosoids, our results indicate that the 3 tribes Mimoseae (Luckow et al., 2003), Ingeae, and Acacieae (including the genera *Acacia* s.s., *Senegalia*, and *Vachellia*) are all nonmonophyletic (Figure 2), generally consistent with previous studies (Chappill and Maslin, 1995; Käss and Wink, 1996; Luckow et al., 2003; Pertea et al., 2003; Kyalangalilwa et al., 2013; LPWG, 2017; Zhang et al., 2020b; Koenen et al., 2020). The analyses here yielded a largely consistent and well-resolved topology that divides the sampled mimosoids into 17 clades. Among these, Mimoseae (with 11 genera, a quarter of the mimosoid genera) are resolved into a grade of 5 clades that are successive sisters to the remainder of the mimosoids. Among the 5 Mimoseae clades, the clade that contains *Pentaclethra* and 3 other genera is sister to other mimosoids (Figure 2), whereas these 4 genera were previously in a polytomy with a clade of the remaining mimosoids (LPWG, 2017) (Supplemental Figure 1F). See the Supplemental text for a more detailed description and discussion of relationships among the other mimosoid taxa in comparison with previous studies (Herendeen et al., 2003; Luckow et al., 2003; Bouchenak-Khelladi et al., 2010; Kyalangalilwa et al., 2013; LPWG, 2017; Zhang et al., 2020b).

Highly resolved relationships in Papilionoideae

In the largest subfamily, Papilionoideae, we sampled 307 species from 222 genera (Figures 3, 4, and 5) and found consistent and highly supported relationships among tribes and equivalent clades (Figure 3). Three lineages, the ADA clade (with monophyletic tribes Angylocalyceae, Dipterygeae, and Amburaneae), Swartzieae, and Cladrastideae, are successive SPSLs to the remaining Papilionoideae (the 50-kb inversion clade, Figure 3). Previously, the relationships among these three lineages and other Papilionoideae were inconsistent; either the ADA clade and Swartzieae together (50%–60% BS) (Wojciechowski et al., 2004), or the ADA clade (Cardoso et al., 2012, 2013), or Swartzieae were sister to the remaining Papilionoideae (LPWG, 2017; Zhang et al., 2020b).

Previously, there was a polytomy within the 50-kb inversion clade, with unresolved relationships among 10 lineages (Wojciechowski et al., 2004; Cardoso et al., 2012, 2013), i.e., *Amphimas*, *Aldina*, Exostyleae, vataireoid, *Dermatophyllum*, dalbergioids s.l., genistoids s.l., the *Andira* clade, Baphieae, and the NPAAA clade. In the present phylogeny with samples from seven of these lineages, Exostyleae and vataireoid are placed as successive sisters to the remaining 50-kb inversion clade, which forms two large clades. One contains nine tribes divided into two subclades of dalbergioids s.l. and genistoids s.l., which contain two and seven tribes, respectively (Figures 1 and 3). The other is resolved into the *Andira* clade, Baphieae, and the NPAAA clade (Figures 1 and 4; see the Supplemental text for additional information on relationships among these lineages).

In Papilionoideae, a large clade called the NPAAA clade (also known as the CA clade) (Figures 1, 4, and 5) is represented here by 207 species (over 2/3 of sampled papilionoids) and is monophyletic with relatively high support (at least 4 trees with BS > 80%); in addition, *Andira* and Baphieae are successive sisters to the NPAAA clade (Figure 4). Previous phylogenies using plastid genes also supported Baphieae as the sister to the NPAAA clade (Cardoso et al., 2012; Cardoso et al., 2013; LPWG, 2017; Zhang et al., 2020b), whereas the position of *Andira* was not resolved relative to other SPSLs of the remaining Papilionoideae (Cardoso et al., 2012, 2013). Within the NPAAA clade, our results support Hypocalypteae as the sister to the remaining taxa (at least 6 trees with BS > 80%), whereas the Mirbelioids and Hologalegina clades are successive sisters to Indigofereae + millettioids (Figures 1, 4, and 5). Within Hologalegina, Robinieae and Sesbanieae + Loteae are successive sisters to the IR-lacking clade (inverted repeat-lacking clade), which is further resolved into five subclades (Figure 4). Recently, a phylogeny using whole plastomes (with 24 species for NPAAA) (Zhang et al., 2020b) strongly supported the same topology. Additional information on the relationships within the large clades of the NPAAA clade is included in the Supplemental text (Figure 4) and the following section (Figure 5).

The millettioids/phaseoloids (referred to as millettioids hereafter for convenience) are a large clade (Figure 5) that contains many economically important legumes such as *Glycine max* (soybean), *Phaseolus vulgaris* (common bean), *Vigna* spp. (cowpea, mung bean), and *Cajanus cajan* (pigeon pea), as well as some important forage plants such as *Lespedeza* and *Desmodium*. However, the relationships among millettioids, especially the polyphyletic tribe Phaseoleae, still include a number of uncertainties (Stefanović et al., 2009; LPWG, 2017). Our results maximally support the tribe Indigofereae as monophyletic and sister to the millettioids (Figure 4), consistent with strongly supported hypotheses proposed in previous studies (Hu et al., 2000; Kajita et al., 2001; Wojciechowski, 2003; LPWG, 2017; Zhang et al., 2020b). Within the millettioids (Figure 5), the genera *Aganope*, *Austrosteenisia*, and *Dalbergiella* are successive sisters to the remaining millettioids with high support (at least four trees with BS > 90%). The next diverging clade contains three tribes, and Abreae is sister to Diocleae + Millettieae, consistent with previous studies (de Queiroz et al., 2015; LPWG, 2017).

The remaining millettioids include members of three tribes, Phaseoleae (including Clitoriinae), Desmodieae, and Psoraleeae (Figure 5). Here, Phaseoleae is paraphyletic, with Desmodieae and Psoraleeae nesting within Phaseoleae, as reported previously (Kajita et al., 2001; Egan et al., 2016; LPWG, 2017). Specifically, the clade containing Phaseoleae, Desmodieae, and Psoraleeae is resolved into 11 lineages (I to XI), comprising *Platycyamus regnellii* (here called "lineage X") plus 10 monophyletic clades (Figure 5). Among the 11 lineages, lineage I contains the tribe Psoraleae and the subtribe Glycininae of Phaseoleae, with members of Glycininae forming a grade of 8 clades, including 1 clade of *Glycine max* (soybean) and its close relatives (Figure 5). Lineage III is the clade of the subtribe Phaseolinae, which includes several cultivated species such as common bean, lima bean, and mung bean (Figure 5). In lineage IV, two genera of Phaseoleae, *Craspedolobium* and *Haymondia*, are successive sisters to the monophyletic tribe Desmodieae with 12 genera sampled here, whereas lineage V is the monophyletic subtribe Cajaninae (including pigeon pea) of Phaseoleae. Taken together, the full Fabaceae phylogeny provides crucial support

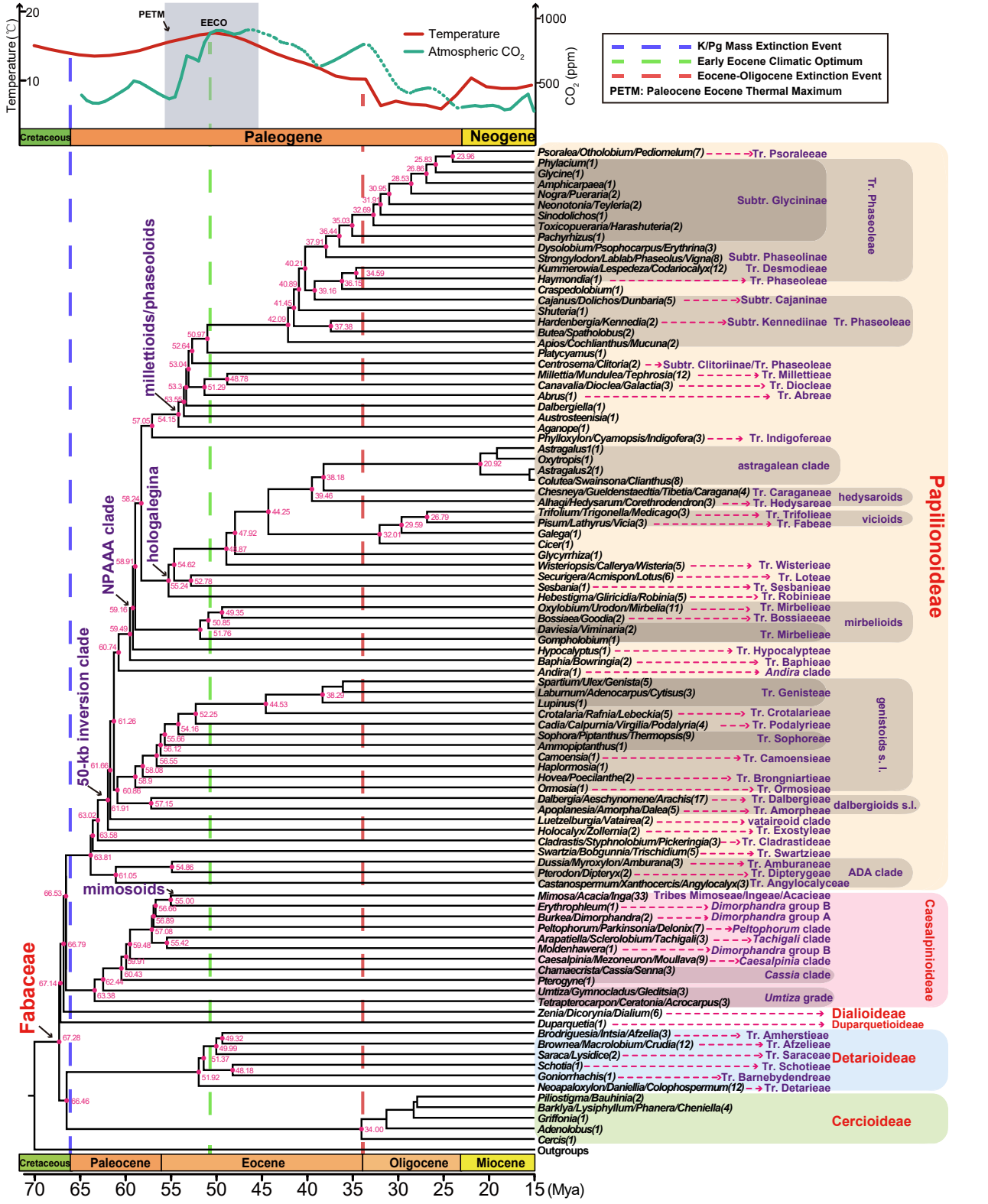

Figure 6. Divergence times of major Fabaceae clades.

This summary tree is a simplified version with tips collapsed as noted in the tip names (as in Figure 1). Numbers after the tips show the total number of genera sampled for each collapsed tip. Red numbers next to nodes indicate the median ages of the divergence times. The top panel shows the red curve of global temperature change in Earth's history based on data compiled from previous studies (Veizer et al., 2000; Zachos et al., 2008). The green curve

(legend continued on next page)

for the revision of a number of taxa and reveals several differences in phylogenetic placement between plastid and nuclear trees.

Divergence of Fabaceae near the boundary of the late Cretaceous and Paleocene

The extensive taxonomic and molecular sampling here enabled the estimate of origins and divergence times of multiple Fabaceae lineages, particularly at intrasubfamilial levels. We used the model tree summarized from 7 coalescent trees (Figures 2, 3, 4, and 5, and Supplemental Figure 3) with branch lengths from the 131-gene concatenated supermatrix of 479 species (463 legumes and 16 other eudicots) as input to estimate divergence times by a penalized likelihood (PL) method implemented in treePL (Smith and O'Meara, 2012). We used twenty-three fossil-based age constraints, along with the tricolpate pollen fossil, to specify the age of crown eudicots (Supplemental Table 5) (Doyle et al., 1977; Lavin et al., 2005; Bruneau et al., 2008; Silveira et al., 2016). The estimated mean ages of major clades and tribes are shown in Figure 6, with their 95% confidence interval estimates presented in Supplemental Table 6 and additional age estimates in Supplemental Figure 6. We estimated the legume crown age to be ~67.28 mya (million years ago), close to the K/Pg boundary (Figure 6) and in line with the estimate of Koenen et al. (2020b), but slightly older than the previous estimate of ~64 mya based on three plastid sequences (Bruneau et al., 2008). Shortly afterward, six subfamilies diverged rapidly within ~0.82 my, near 67 mya. Most tribes of Caesalpinioideae and Papilionoideae diverged from the late Paleocene to the early Eocene (from ~64 to ~50 mya), including the two species-rich clades (the millettioids and Hologalegina) (~58.24 mya). The NPAAA clade, the largest subgroup of Papilionoideae that includes the majority of cultivated legumes, diverged from the tribe Baphieae at ~59.49 mya. All Detarioideae tribes and some Papilionoideae tribes (such as Desmodieae and several lineages of Phaseoleae) originated during the middle to late Eocene (~50–35 mya) (Figure 6). Previously, crown groups represented by extant Caesalpinioideae, mimosoids, and Papilionoideae, as well as the tribes Cercideae (now subfamily Cercidoideae) and Detarieae, were estimated to have emerged between 34 and 63.7 mya, a range similar to that in our study (Lavin et al., 2005; Bruneau et al., 2008). These findings all point to a rapid family-wide diversification of Fabaceae.

Numerous WGD events across Fabaceae supported by clusters of GDs

WGD or polyploidy has been supported by the analyses of chromosome numbers both in early Fabaceae history and in specific lineages (for a review, see Doyle, 2012). In the past decade, nearly 40 legume genomes have been sequenced (Egan and Vatanparast, 2019) (https://peanutbase.org/data/public/), facilitating the detection of molecular evidence for WGD in Fabaceae using genomic synteny analyses (Cannon et al., 2006; Schmutz et al., 2010; Hane et al., 2017; Wang et al., 2017; Stai et al., 2019; Zhuang et al., 2019) or intragenomic gene cluster analyses (e.g., Egan and Doyle, 2010). In addition, phylotranscriptomics or comparative genomics has been used for the phylogenetic detection of GD clusters in the ancestors of Fabaceae and several subfamilies (Cannon et al., 2015; Leebens-Mack et al., 2019; Stai et al., 2019; Koenen et al., 2020b), as performed previously for other plant groups (e.g., Jiao et al., 2011; Huang et al., 2016b; Ren et al., 2018; Zhang et al., 2020c). However, previous legume studies either lacked key lineages (such as the subfamily Duparquetioideae; Koenen et al., 2020b) or sampled limited numbers of legumes.

Here, the Fabaceae phylogeny and large sequence datasets from hundreds of legumes, including greater representation of the SPSL of most of papilionoids, provide an excellent opportunity to uncover additional WGDs and to more precisely place the previously reported WGDs onto the Fabaceae phylogeny. We examined a total of 246 822 gene trees from 13 groups of legumes (see Methods; Supplemental Figures 8–21) and identified multiple GD clusters shared by at least 2 species and considered to support WGD or WGT when either one of the 2 criteria was met: (1) GD number >450, GD ratio >4.5%, and percent of ABAB type >50%; or (2) GD number >1000, GD ratio >5%, and percent of ABAB type >20%. Our results support 27 candidate WGDs and 1 WGT (IDs of WGDs/WGT given as 1–28; see Figures 1, 2, 3, 4, and 5, Supplemental Figure 3, and Supplemental Table 7 for positions and Supplemental information), as well as numerous other clusters of GD bursts in Fabaceae (see Supplemental Figures 8–21). WGD1 is shared by all legumes (Figure 1). In addition, there are 2, 2, 1, 6, and 16 WGD events, respectively, in Cercidoideae, Detarioideae, Dialioideae, Caesalpinioideae, and Papilionoideae among the taxa sampled in our study. The subfamily Duparquetioideae contains a single species and did not meet the requirement for at least two species; it was not considered for a possible lineage-specific WGD.

Nine genera and 14 species were sampled in Cercidoideae, and 1 WGD was strongly supported in the most recent common ancestor (MRCA) of Cercidoideae (WGD2; Figure 1 and Supplemental Figure 3), consistent with previous analyses of one species each from *Cercis* and *Bauhinia* (Cannon et al., 2015; Leebens-Mack et al., 2019; Koenen et al., 2020b). However, a recent comparative genomic analysis of *Cercis*, *Bauhinia*, and other legumes argued for the lack of polyploidy in *Cercis* (Stai et al., 2019), in part due to its relatively small number of chromosomes, although these results were still consistent with an early Cercidoideae WGD shared by other Cercidoideae. Another WGD event (WGD27; Supplemental Figure 3) was shared by 4 genera (*Barklya*, *Cheniella*, *Phanera*, and *Lysiphyllum*). One WGD was detected at the MRCA of Detarioideae, with strong support from analyses of sequences from 32 species in 29 genera (WGD3; Figure 1 and Supplemental Figure 3). This WGD was consistent with previous studies of sequences from 1 to 4 species (Cannon et al., 2015; Leebens-Mack et al., 2019; Koenen et al., 2020b) but received support from multiple tribes here (see next section

indicates ancient atmospheric CO_2 levels compiled from a previous study (Beerling and Royer, 2011); the portions with dashed lines indicate significant decreases. The bottom shows the geological timescale in million years ago (mya). Vertical dashed lines highlight important geological events (blue, K/Pg boundary; green, Early Eocene Climatic Optimum; red, Eocene-Oligocene extinction event).

for further discussion). Another WGD in Detarioideae was detected at the MRCA of *Lysidice* (WGD2; Supplemental Figure 3). Also, a WGD (WGD4; Figure 1 and Supplemental Figure 3) was strongly supported at the MRCA of Dialioideae based on the three genera sampled here; it had the highest GD number and ratio of all polyploidy events reported here (Figure 2 and Supplemental Figure 3 and Supplemental Table 7).

Six WGDs were detected in Caesalpinioideae (Figure 2 and Supplemental Table 7), with one at the origin of the subfamily (WGD5), consistent with previous analyses of 5 or 7 species (Cannon et al., 2015; Leebens-Mack et al., 2019). Among the others, WGD7 was shared by the three genera *Umtiza*, *Gymnocladus*, and *Gleditsia*, and WGD8 was shared by *Cassia* species. In the *Caesalpinia* clade, WGD9 was shared by the 4 genera *Mezoneuron*, *Pterolobium*, *Biancaea*, and *Moullava*. Two WGD events were identified in the mimosoids, one shared by two *Prosopis* species (WGD10) and the other shared by *Samanea*, *Senegalia*, and *Mimosa bimucronata* (section *Batocaulon* DC.) (WGD11). In addition, *Prosopis* (2n = 28, 52, 56) was reported to include polyploid species based on analyses of chromosome number (Doyle, 2012). However, flow cytometry analysis and literature review indicated that the majority of *Prosopis* species, including *P. glandulosa* and *P. velutina* analyzed here, are diploid (Trenchard et al., 2008). Such diploid legumes (see below for additional examples) that have experienced WGD but have relatively low chromosome numbers can be considered cryptic polyploids or as having been paleopolyploids, as discussed previously (Doyle 2012).

We detected 17 WGD/WGT events in Papilionoideae (Figures 3, 4, and 5), 1 of which was shared by the entire subfamily (WGD6) (Figure 3), as has been reported previously (Cannon et al., 2015; Leebens-Mack et al., 2019; Koenen et al., 2020b) (see also next section). Another WGD is present in Dalbergieae (Figure 3) and shared by two *Pterocarpus* species (WGD12). In genistoids s.l., a WGD is shared by *Thermopsis* and *Piptanthus* in Sophoreae (WGD13), and another is present in Crotalarieae and shared by *Aspalathus*, *Cyclopia*, *Lebeckia*, and *Rafnia* (WGD14) (Figure 3). In addition, a WGT event was identified at the MRCA of Genisteae (WGT15) (Figure 3), with further support from triplicate syntenic blocks in the genome of lupin (*Lupinus angustifolius*) (Supplemental Figure 22; see below). Members of *Thermopsis* (2n = 18, 36) and *Pterocarpus* (2n = 22, 24, 44), as well as the entire Genisteae tribe, were proposed to be polyploid based on observations that members of these groups exhibit extensive variations in chromosome number and that some have large numbers of chromosomes (Doyle, 2012).

In the NPAAA clade (Figures 4 and 5), a WGD was detected at the MRCA of Mirbelieae (WGD16; Figure 4). Also, four WGDs were identified in the Hologalegina clade (Figure 4): WGD17 at the MRCA of Fabeae, WGD18 shared by *Corethrodendron* and *Hedysarum*, WGD19 in *Tibetia*, and WGD20 in *Caragana* (Supplemental Figure 16). Among the five *Caragana* species sampled here, *C. sinica* was reported to be a triploid, whereas the others are diploids (Zhang et al., 2009). We further tested the MRCA of these 5 *Caragana* species for paleopolyploidy, with *Tibetia* and *Astragalus* species as outgroups. Tree reconciliation identified 682 GDs in the MRCA of the sampled *Caragana* taxa (Supplemental Figure 17). In addition, Ks plots of paralogs in each *Caragana* species corresponding to the GDs mapped to the MRCA of the five *Caragana* species showed a peak (Ks at ~0.25) (Supplemental Figure 17). Both the GD cluster and the Ks peaks further support paleopolyploidy of the MRCA of the sampled *Caragana*, distinct from the polyploidy of some species in this genus supported by chromosome numbers (Zhang et al., 2009). Furthermore, six WGDs were identified in millettioids/phaseoloids (Figure 5): WGD21 shared by *Canavalia* in Diocleae, WGD22 in *Grona* in Desmodieae, WGD23 at the MRCA of *Neonotonia* and *Teyleria*, WGD24 in *Teyleria*, WGD25 at the MRCA of *Glycine*, and WGD26 at the MRCA of Psoraleeae.

The WGD/WGT events shared by species with sequenced genomes (WGD1, WGD2, WGD5, WGD6, and WGT15) were further examined by synteny analysis, and GDs in clusters that supported WGDs matched paralog pairs located in syntenic blocks (see Methods and Supplemental Table 8). Our analyses uncovered genomic evidence for WGD1 shared by Fabaceae in the representative legume genomes of *Medicago truncatula* (Young et al., 2011) (20/895, indicating the number of syntenic paralog pairs/total GD number in the cluster that supports the WGD), *Arachis duranensis* (Bertioli et al., 2016) (21/344), *Phaseolus vulgaris* (Schmutz et al., 2014) (20/320), *Glycine max* (Schmutz et al., 2010) (96/1137), and *Cercis canadensis* (Griesmann et al., 2018) (3/730) (Supplemental Table 8). Paralog pairs were also found (64/3179) in syntenic regions of the *Cercis canadensis* genome supporting WGD2 shared by members of Cercidoideae (Supplemental Table 8). For WGD5 shared by members of Caesalpinioideae, paralog pairs were also found in syntenic regions in the genomes of *Mimosa pudica* (Griesmann et al., 2018) (17/231), *Chamaecrista fasciculata* (Griesmann et al., 2018) (20/265), and *Prosopis alba* (https://www.ncbi.nlm.nih.gov/nuccore/SMJV00000000.1/) (10/253). However, the numbers of paralog pairs placed at the MRCA of Caesalpinioideae are relatively small (Supplemental Table 8), possibly due to the relatively low quality of sequenced genomes with short contigs in this subfamily. For WGD6 shared by members of Papilionoideae, paralog pairs were found in syntenic regions in the genomes of *Glycine max* (1621/4696), *Medicago truncatula* (329/1485), *Phaseolus vulgaris* (387/1249), and *Arachis duranensis* (305/1910). As mentioned above, paralogs (355/1322) supporting WGT15 shared by the members of Genisteae were detected in syntenic regions in the lupin (*Lupinus angustifolius*) genome (Hane et al., 2017) (Supplemental Table 8). Specifically, several triplicate syntenic regions detected from a dot plot of syntenic paralogous genes in the longest 59 scaffolds of the lupin genome (Supplemental Figure 22) (Hane et al., 2017) served as evidence for the WGT at the MRCA of Genisteae. Moreover, previous comparative genomic analyses of *Lupinus angustifolius* with *Lotus japonicus*, *Medicago truncatula*, *Phaseolus vulgaris*, and *Glycine max* have uncovered syntenic evidence of a WGT in *Lupinus*, and Ks analyses place its age at ~25 mya in genistoids (Hane et al., 2017), a date that predates the radiation of *Lupinus* (Hughes and Eastwood, 2006).

As mentioned above, polyploidy in legumes has been investigated using the cytological analyses of chromosome numbers (Doyle, 2012). Specifically, polyploidy during early legume history was examined based on either Cercidoideae

(Cercideae) or Detarioideae (Detarieae) as the sister to all other legumes; the former scenario possibly suggests a lower ancestral legume chromosome number ($x = n = 7$) than the latter ($x = n = 12$). Members of two other subfamilies, Dialioideae and Caesalpinioideae, also have relatively high base chromosome numbers ($x = 14$). The situation for Papilionoideae is more complex, with a high number of $x = 28$ starting at the ancestor of Genistoids, Dalbergioids, and many other papilionoids, or possibly earlier at the ancestor of the entire Papilionoideae. These scenarios for individual subfamilies are consistent with the supported WGD events for each of the five multi-species subfamilies (Figure 1). In addition, the grouping of Cercidoideae and Detarioideae together as sister to other legumes supported by this and other studies (Figure 1) (Koenen et al., 2020a) favors a higher ancestral legume chromosome number, consistent with the proposed WGD at the MRCA of Fabaceae (see next section for a detailed discussion).

Moreover, Doyle (2012) reported a survey of nearly 400 genera for polyploidy based on chromosome numbers, and about a quarter of those genera had at least some polyploid taxa. For example, *Prosopis* (2*n* = 28, 52, 56), *Thermopsis* (2*n* = 18, 36), many members of Genisteae, *Pterocarpus* (2*n* = 22, 24, 44), several Mirbelioids, *Caragana* (2*n* = 16, 24, 32, 48), *Glycine* (2*n* = 38, 40), and *Teyleria* (2*n* = 44) are associated with the WGDs supported here (Figures 2, 3, 4, and 5). However, as these genera usually also have diploid species, their polyploids may not be directly linked to the WGDs that are shared by two or more genera, or even to those shared by two or more species in the same genus such as those in *Caragana* (see above for a discussion of WGD20). *Prosopis* (WGD10) is a good example—this WGD event is supported by a relatively high percentage of 21.1% GD (Supplemental Table 7). This is somewhat surprising given that a comprehensive flow cytometry analysis and literature review of ploidy levels across the genus overwhelmingly supported the majority of species as diploids (Trenchard et al., 2008). If such diploids with relatively low chromosome numbers have indeed experienced WGD, they would be referred to as cryptic polyploids or paleopolyploids, as discussed previously (Doyle, 2012).

To gain clues about gene functional categories affected by WGDs, we examined the gene ontology (GO) of retained gene duplicates from the ancestral Fabaceae WGD. The retained duplicates are enriched for several GO categories (Supplemental Figure 23), primarily response to external biotic stimulus, defense response to bacterium, transmembrane transporter activity, and plasma membrane and protein metabolic processes. Increased gene copies with some of these annotated functions may have played a role in enhanced interaction with rhizobial bacteria and increased the synthesis of proteins, including seed storage proteins, during the evolutionary history of Fabaceae. In addition, previous comparative genomics studies suggested that the ancestral papilionoid polyploidy event led to enhanced root nodule symbiosis in this largest legume subfamily (Young et al., 2011; Li et al., 2013), consistent with a WGD event in the MRCA of Papilionoideae that is supported by the phylogenomic analysis here and by previous studies (Cannon et al., 2015; Leebens-Mack et al., 2019; Stai et al., 2019; Koenen et al., 2020b).

Examination of alternative hypotheses for WGDs in early Fabaceae history

Analyses of legume genomes have uncovered syntenic evidence for WGD in *Glycine max*, *Medicago truncatula*, *Lotus japonicus*, *Phaseolus vulgaris*, and *Lupinus angustifolius*, including WGDs shared by divergent lineages of Papilionoideae (Cannon et al., 2006; Schmutz et al., 2010; Hane et al., 2017; Wang et al., 2017; Stai et al., 2019; Zhuang et al., 2019). In addition, comparative genomics and phylotranscriptomics support WGDs during early Fabaceae history. Specifically, phylotranscriptomic studies using sequences from 20 (Cannon et al., 2015) or 26 legumes (Leebens-Mack et al., 2019) provided support for the ancestral papilionoid WGD with the sampling of 12 (or 16) species, including *Xanthocercis* and *Cladrastis*, which represent 2 of the SPSLs (the ADA clade and Clasdrastideae) of most Papilionoideae (see Figure 3 for placement of these lineages). In addition, evidence was obtained for WGDs in early Cercidoideae (one species in each of *Cercis* and *Bauhinia*), Detarioideae (*Copaifera officinalis*), and early Caesalpinioideae (5 species, one in each of *Chamaecrista*, *Senna*, *Desmanthus*, *Gymnocladus*, and *Gleditsia*, or 7 species). However, the Ks peak for *Cercis* paralogs provided weak support for the WGD in early Cercidoideae (Cannon et al., 2015). Indeed, a study of *Cercis canadensis*, *Bauhinia tomentosa*, and other legumes argued for the lack of polyploidy in *Cercis* (Stai et al., 2019), as supported by its relatively small chromosome number, although these results are still consistent with an early WGD shared by other Cercidoideae. The analyses here support WGDs in the early histories of Papilionoideae, Caesalpinioideae, and Cercidoideae with sequences from more taxa, including a greater representation of the SPSL of most of papilionoids and 14 species in 9 Cercidoideae genera (Figures 2, 3, 4, and 5 and Supplemental Figure 3). Moreover, the early Detarioideae WGD is supported by the analyses of sequences from 32 species in 29 genera (Supplemental Figure 3). WGD4 at the base of Dialioideae is supported by the highest number of GDs detected and a 39.7% GD ratio, the highest of all polyploidy events here (Supplemental Table 7). Compared with previous studies (Cannon et al., 2015; Koenen et al., 2020b) with only two representatives (*Bauhinia tomentosa* and *Cercis canadensis*), the inclusion of many species here appeared to permit the detection of more paralogs, providing greater support for these WGD events, particularly those involving Cercidoideae, wherein >10% of GD supported this event (Supplemental Table 7).

On the other hand, previous analyses of legume genomic and transcriptomic datasets did not report the WGD event detected here at the MRCA of Fabaceae (Figure 1 and Supplemental Figure 21), although they may have detected low or insignificant signals (Cannon et al., 2015; Leebens-Mack et al., 2019; Stai et al., 2019; Koenen et al., 2020b). Koenen et al. (2020b) reported evidence for possible polyploidy associated with the early evolutionary history of legumes, including one at the MRCA of Fabaceae (Figure 1 of Koenen et al. [2020b]); however, they concluded that allopolyploidy for a subset of Fabaceae subfamilies, such as for the MRCA of Papilionoideae, Caesalpinioideae, and Dialioideae, was more probable. Such

hypotheses of allopolyploidy were based on the analyses of gene family trees with genes from one genome (*Medicago truncatula*/Papilionoideae) and 7 transcriptomes (*Bauhinia tomentosa*/Cercidoideae, *Anthonotha fragrans*/Detarioideae, *Zenia insignis*/Dialioideae, and 4 Caesalpinioideae species, *Albizia julibrissin*, *Entada abyssinica*, *Inga spectabilis*, and *Microlobius foetidus*). These allopolyploidy hypotheses include the second parental lineage as the ancestors of Cercidoideae, Detarioideae, or both subfamilies.

To further compare Fabaceae ancestral WGD with the allopolyploidy hypotheses for a subset of subfamilies, we examined the 5088-gene family trees with sequences from 70 species, including 65 legumes (Supplemental Figure 21), with a GD cluster (gene duplication node BS $\geq$ 70 and species with duplicates $\geq$2) that supported the WGD event (WGD1) at the MRCA of Fabaceae. We reasoned that gene families with two paralogs from the same subfamily (Supplemental Figure 58; sps.duplication) would support the hypothesis that WGD1 occurred at the MRCA of Fabaceae, whereas gene families with only one paralog for a subfamily (Supplemental Figure 58; sps.sister and sps.other) would support the corresponding subfamily being the second parental lineage of an allopolyploid. Therefore, for each species, we counted the proportion of gene families that belonged to one of three types (Supplemental Figure 58). The great majority of genes in each species belonged to gene families with two paralogous clades for the same subfamily ($p < 0.001$), supporting the WGD at the MRCA of Fabaceae rather than at the MRCA of Papilionoideae and Caesalpinioideae (or also Dialioideae).

Following a polyploidy event and subsequent species divergence, different paralogs may be lost in different species lineages such as legume subfamilies (Egan and Doyle, 2010). According to the Fabaceae phylogeny supported by analyses here and Koenen et al. (2020a, 2020b), Cercidoideae and Detarioideae form a clade that is sister to the clade of other legumes, including Papilionoideae, Caesalpinioideae, and Dialioideae. Therefore, gene families with two paralogous clades, each with genes from both (1) Cercidoideae and/or Detarioideae and (2) any of Papilionoideae, Caesalpinioideae, and Dialioideae, support the polyploidy at the MRCA of Fabaceae (see Supplemental Figure 59A for several examples; see also Supplemental Table 13 for details of 140 such topologies). On the other hand, gene families with some other patterns of gene loss (Supplemental Figure 59B) (such as when one of the Cercidoideae or Detarioideae has a paralog and the other subfamily has none) support one of the hypotheses proposed by Koenen et al. (2020b). From gene families of the 70-species analysis (65 legumes; Supplemental Figure 21), 1015 GDs (gene set1; Supplemental Figure 59A) were identified when we required a minimum of 50% BS and at least 4 species with 2 paralogs, whereas 663 GDs (gene set2) were identified when we required at least 70% BS and 5 species with 2 paralogs. In gene set1 and gene set2, 93.30% and 91.55% GDs, respectively, show topologies that support the Fabaceae ancestral polyploidy (Supplemental Figure 59A), and less than 3% show topologies that support the allopolyploidy hypotheses (Supplemental Figure 59B) for the MRCA of Papilionoideae and Caesalpinioideae (or with Dialioideae).

Fabaceae phylogeny, polyploidization, and N_2 fixation

With the loss of distinct paralogs in different taxa, one might expect that the analysis of sequences from more species would increase the chance of detecting both paralogs from a polyploid event. In addition, transcriptome datasets are likely to have less complete gene content than genomic datasets. As our analyses included more legumes (65) and a relatively large number (24) of sequenced genomes compared with those of Koenen et al. (2020a, 2020b) (8 species with 1 genome), we speculated that differences between the results of the two studies might be due to differences in the number of sampled species. To test this possibility, we constructed gene trees using sequences from 16 species (14 legumes, reduced from the 65 legumes in the above-mentioned analysis), including 5 genomes (2 Papilionoideae, 2 Caesalpinioideae, and 1 Cercidoideae) and 11 transcriptomes with relatively high N50 of assembled unigenes and/or completeness with reference to BUSCO conserved genes (3 Dialioideae, 2 Cercidoideae, 4 Detarioideae, and 2 outgroups; see Supplemental Table 2 for assembly information). From this analysis, 933 GDs (gene set3; Supplemental Figure 59A) were detected at the MRCA of Fabaceae with at least 50% BS and 2 species with two paralogs; among the 933 gene trees, >81% had topologies that supported WGD1 at the MRCA of Fabaceae. A fourth gene set was obtained from another analysis of 48 species (Supplemental Figure 61), and >78% of the trees had subfamily distributions that supported the Fabaceae ancestral polyploidy (Supplemental Figure 59A).

Furthermore, we used gene set3 (from 16 species) to analyze possible allopolyploidy using the GRAMPA program (Supplemental Figure 60). Three allopolyploidy models (Supplemental Figure 60B, a–c) had reconciliation scores (lower scores imply better support from the data) ranging from 11 502 to 11 868, lower than the score for polyploidy (11 964) at the MRCA of Fabaceae (Supplemental Figure 60), although the differences were 462 or less. However, our allopolyploidy models (Supplemental Figure 60B, a–c) all proposed that the MRCA of Cercidoideae + Detarioideae was a hybrid involving a second parental lineage as the ancestor of one or more of the subfamilies Papilionoideae, Caesalpinioideae, or Dialioideae. These models differed from those proposed by Koenen et al. (2020b), i.e. that the MRCA of Papilionoideae and Caesalpinioideae, or the MRCA of Papilionoideae, Caesalpinioideae, and Dialioideae, was a hybrid and that a second parental lineage was either Cercidoideae, or Detarioideae, or the MCRA of Cercidoideae and Detarioideae. As the 16-species analysis here contained more Cercidoideae and Detarioideae species, including a *Cercis* genome, than Papilionoideae or Caesalpinioideae species, the results suggest that the small species number (two each) of Papilionoideae and Caesalpinioideae led to a failure to detect one or both paralogs in these subfamilies for many gene pairs. It is therefore likely that when the number of sampled species in a subfamily is small, there is a greater chance that the failure to detect both paralogs in gene pairs in a subfamily may yield a misleading signal for allopolyploidy in the MRCA of a subset of Fabaceae subfamilies. Specifically, the lineage(s) with fewer sampled species may be designated as the second parent of the clade with subfamilies that have more sampled species (and thus more detected duplicates). In short, our analyses using a much larger number of species, with multiple representatives for each of five legume

subfamilies, yielded substantial support for a polyploid (tetraploid) at the MRCA of Fabaceae.

Ancestral state reconstruction of rhizobial nodulation symbiosis in Fabaceae

Nitrogen-fixing Fabaceae species with rhizobial nodules have been identified in the two largest subfamilies, Papilionoideae and Caesalpinioideae (Doyle, 2016); however, the early histories of rhizobial nodulation in Fabaceae have been unclear because of the uncertain relationships among the SPSLs of other members of these subfamilies (Supplemental Figure 1). Using the well-resolved Fabaceae phylogeny as a reference, we performed character reconstruction analyses of rhizobial nodules for nitrogen fixation (Supplemental Figures 24 and 25) using maximum parsimony and ML-based methods (Supplemental Table 9). The results suggest that the MRCA of Fabaceae probably did not produce rhizobial nodules, nor did each of the four smaller subfamilies without known rhizobial nodulation (Supplemental Figure 24), as proposed previously (Doyle, 1998; Werner et al., 2014; Li et al., 2015). The ancestor of Caesalpinioideae is proposed here to be non-nodulating, as supported by the non-nodulation state of three successive sisters (the *Umtiza* grade and *Pterogyne*) of the remaining Caesalpinioideae. Within Caesalpinioideae, rhizobial nodulation is proposed for *Chamaecrista* (in the *Cassia* clade); however, the rhizobial nodulation state of the backbone is equivocal after the divergence of the *Cassia* clade and until the mimosoid clade (e.g., Supplemental Figure 24A). The ancestor of Papilionoideae and four additional nodes along the backbone of Papilionoideae are proposed to be non-nodulating, as supported by the non-nodulation state of the ADA clade, the tribes Clasdrastideae and Exostyleae, and the vataireoid clade, in a grade of five successive sisters of other Papilionoideae (Supplemental Figures 24 and 25). The ancestor of the Swartzieae is proposed to be nodulating, as is *Vatairea*. After the divergence of the vataireoid clade, the common ancestor of the remaining Papilionoideae is also proposed to be nodulating. Furthermore, losses of nodulation are proposed for *Nissolia*, *Bowringia*, and *Chesneya*.

More recently, there has been increasing acceptance of a hypothesis in which a single origin of (or a predisposition for) nitrogen-fixing nodulation in the nitrogen-fixing clade of four orders (Fabales, Fagales, Rosales, and Cucurbitales) was followed by massive, multiple losses, given the complexity of the nodulation process that requires many gene functions, the well-supported single clade that contains all nitrogen-fixing taxa, and the expression results of multiple genes in both rhizobial- and actinorhizal-nodulating species (Soltis et al., 1995; Sprent, 2009; Doyle, 2011; van Velzen et al., 2017, 2018, 2019; Battenberg et al., 2018; Griesmann et al., 2018). This second hypothesis is also supported by molecular analysis of multiple gene families related to rhizobial nodulation for nitrogen fixation.

Evolutionary histories of key nodulation genes

The hypothesis of massive multiple losses of nodulation has been supported by large-scale analysis of nodulation-related genes. In particular, homologs of *LjNIN* (*Lj* = *Lotus japonicus*), which encodes a regulator of nodulation-related gene expression (e.g., Suzuki et al., 2013), were found to be lost (Griesmann et al., 2018) in multiple non-nodulating species of the nitrogen-fixing clade (Fabales, Fagales, Rosales, and Cucurbitales) (Soltis et al., 1995; Sprent, 2009; Doyle, 2011). In addition, legumes form rhizobial nodules that are distinct from the actinorhizal nodules in most of the non-legume nitrogen-fixing species distributed in three orders. Nitrogen-fixing actinorhizal nodules may be ancestral, and there may have been one or more switch(es) from actinorhizal nodules to rhizobial nodules (van Velzen et al., 2019), such as that suggested for *Parasponia* (van Velzen et al., 2017, 2018).

To examine the evolution of legume rhizobial nodulation at the molecular level, we performed phylogenetic analyses of homologs from 30 gene families in 28 public genomic datasets, including those of 19 legumes (Supplemental Table 10). Each of these gene families includes genes with known functions in rhizobial nodulation for symbiotic nitrogen fixation (SNF), such as *LjNIN*, *MtRPG*, *LjNFR1/MtLYK3*, *LjLNP*, and *GmAGO5* (*Mt* = *Medicago truncatula*; *Gm* = *Glycine max*; *Lj* = *Lotus japonicus*) (Griesmann et al., 2018; Roy et al., 2020) (Supplemental Figure 26) (see Supplemental Table 11 for information on these and other genes). The amino acid sequences encoded by the functionally studied legume genes were used as queries to search 28 high-quality angiosperm genomes (Supplemental Table 10) using BLASTP with a lenient E-value threshold of less than $1e^{-5}$ and a minimal amino acid sequence identity of 20%. Preliminary phylogenetic analyses identified genes in the same clade that includes all known legume SNF genes; such trees for the 30 key nodulation genes are available publicly (https://github.com/Genomic-docker/Evolution-of-key-nodulaiton-genes). Previously, the presence/absence of *NIN* and 21 other genes were analyzed in 13 legumes and other plants (Griesmann et al., 2018). As we are mainly interested in the evolutionary history of rhizobial nodulation in Fabaceae, we examined the phylogenies of the SNF clades (Supplemental Figures 28–57) to detect possible GD in legumes (Supplemental Figure 26). In addition, the copy numbers of genes in other clades with other possible functions, such as response to nitrate and arbuscular mycorrhizal symbiosis, were also assessed (Supplemental Figure 27).

Our analyses confirmed the previous finding that two non-nodulating legumes, *Cercis canadensis* and *Nissolia schottii* (in Cercidoideae and Papilionoideae, respectively), probably experienced losses of *LjNIN* and *MtRPG* (van Velzen et al., 2017, 2018; Griesmann et al., 2018). A third gene, *LjNFR1/MtLYK3*, was also found to be lost (Supplemental Figures 28–30) in *Cercis canadensis* and *Nissolia schottii*, providing further evidence for the hypothesis of multiple losses of rhizobial nodulation (van Velzen et al., 2017, 2018; Griesmann et al., 2018). The losses were further verified by tBLASTn genome searches to confirm that this result was unlikely to be due to annotation problems. *LjNFR1/MtLYK3* and *LjNIN* are important for the early steps of infection (Supplemental Table 11). The *LjNIN*-related genes were classified previously as *NIN* homologs and *NLPs* (*NIN-like Proteins*) using phylogenetic analysis with a few legumes and *Arabidopsis* (Suzuki et al., 2013). The analyses here with 19 legumes and 9 other angiosperms revealed that these 2 groups probably resulted from the gamma WGT event shared by core eudicots (Jiao et al., 2012). *NLP* homologs were identified in both nodulating and non-nodulating legumes, as well as in non-

legumes, consistent with their functions in nitrate signaling (Konishi and Yanagisawa, 2013). In addition, we found *NIN* homologs in non-legume species, including the non-fixers *Arabidopsis*, tomato, and grape (Supplemental Figures 26 and 28), in agreement with previous findings (van Velzen et al., 2017, 2018; Griesmann et al., 2018) and suggesting that *NIN* homologs have other functions in non-nodulating plants.

Previously, the *LjNIN* region encoding the N-terminal portion of the NIN protein was found to have deletions/mutations, resulting in the loss of its putative ancestral function in nitrate signaling and response. This result led to the hypothesis that functional loss may have allowed for functional innovation of *NIN* in rhizobial nodulation (Suzuki et al., 2013). Our comparison of *NIN* homologs from 19 legumes identified deletions (with possible subsequent mutations) in the N-terminal region of these legume NIN homologs (Supplemental Figure 28). By contrast, the N-terminal regions of NIN homologs in Rosales species and other non-legumes, even those with actinorhizal symbiotic nodules (*Dryas drummondii*, *Datisca glomerata*, and *Casuarina glauca*), appeared to lack this internal deletion found in the legume *NIN* homologs (Clavijo et al., 2015; van Velzen et al., 2019). This finding suggested that the N-terminal deletion in *NIN* homologs probably occurred after the ancestor of legumes diverged from other families in the nitrogen-fixing clade but before the separation of Caesalpinioideae and Papilionoideae. The pattern of *NIN* gene phylogeny and the N-terminal deletion in legume *NIN* homologs support a microsymbiont Switch Hypothesis (SH1, Figure 7) for a switch from actinorhizal to rhizobial nodulation at the MRCA of Fabaceae, although other hypotheses for earlier switches in legume history (such as in the MRCA of Fabales) cannot be ruled out. The proposed SH1 is distinct from the switch proposed for the non-legume rhizobial nodulator *Parasponia* (van Velzen et al., 2019). Similarly, we found that legume *RPG* homologs share specific internal in-frame deletions (Supplemental Figure 29), whereas *RPG* homologs from non-legumes (nodulating and non-nodulating) do not have such deletions, suggesting that the legume *RPG* homologs encode shortened proteins and may have functionally diverged from those of non-legumes.

The *LjNFR1/MtLYK3* genes encode LysM receptor-like kinases, and previous phylogenetic analyses have indicated that eudicot *LjNFR1/MtLYK3* homologs form two major clades (LYK-Ia and LYK-Ib) (Buendia et al., 2018; Rutten et al., 2020), although the duplication of LYK-1b genes in Fabaceae was not examined further. To investigate the evolution of *LYK* genes in Fabaceae, we performed a phylogenetic analysis with multiple homologs from many of the 19 legumes sampled here and recovered both LYK-Ia and LYK-Ib clades in eudicots (Supplemental Figure 30), consistent with previous results based on fewer legume genes (Rutten et al., 2020). In addition, the legume genes in the LYK-Ib clade form three highly supported subclades, each with genes from multiple legumes (Supplemental Figure 30): (1) the NFR1 (SNF) clade, which includes the known SNF genes *LjNRF1/MtLYK3* and genes from other nodulating Papilionoideae and Caesalpinioidae species but lacks genes from *Cercis canadensis* or *Nissolia schottii*; (2) the NFRe1 clade with *MtLYK1* and closely related homologs; and (3) the CERK clade with *MtLYK9*, *LjCERK6* (*LjLYS6*), and closely related homologs. The *LjNRF1/MtLYK3* homolog was also not detected in *Cercis chinensis* using BAC clone sequences and PCR (De Mita et al., 2014). Our finding that *Cercis canadensis* lacks such a gene further suggests that the *LjNRF1/MtLYK3* homolog may have been lost in the ancestor of *Cercis* or earlier. In addition, the lack of an *LjNRF1/MtLYK3* homolog in *Nissolia schottii* suggests a separate loss in Papilionoideae. This result and the losses in *Cercis* together support the idea that the ancestor of this clade gained the specialized function for the recognition of Nod factors, as shown for LjNRF1/MtLYK3 (Bozsoki et al., 2020). Both the NFRe1 and CERK clades have genes from *Cercis canadensis*, *Nissolia schottii*, and other non-nodulating legumes, indicating that these two clades were likely derived from ancestral genes in the MRCA of Fabaceae and further suggesting that the NFR1 clade also originated in the MRCA of Fabaceae. *MtLYK3* (in the NFR1/SNF clade) is required to induce infection thread formation (Limpens et al., 2003), and *LjNFR1* (SNF clade) acts as a master switch that triggers recurrent symbiotic events from *NFRe*-attuned epidermal cells (Murakami et al., 2018). In addition, *NFRe* (in the NFRe clade) also promotes rhizobial nodulation and transcriptional regulation in response to Nod factors (Murakami et al., 2018). The lack of NFR1 homologs in non-nodulating legumes suggests that these genes have undergone multiple losses. By contrast, the NFRe1 and CERK homologs in non-nodulating legumes probably have other functions, although some of them have been shown to contribute to nodulation, including those in *Parasponia* (Cannabaceae, Rosales), the only non-legume that can form rhizobial nodules (Rutten et al., 2020).

Furthermore, analyses of a number of gene families indicated that the SNF clade, with known nodulation-related genes in each of these gene phylogenies, contains homologs from both nitrogen-fixing and non-nitrogen-fixing species. Such genes include *MtLYK1/MtLYK9*, *LYK-II* (*EPR3*), *LjLNP*, *LjAPN1*, *LjFEN1*, *MtVPY*, *MtZPT2*, *MtHMGR1*, *LjnsRING*, *LjRINRK1*, *LjNFR5*, *Leghemoglobin*, *MtNSP2*, *LjCASTOR*, *LjCYCLOPS/MtIPD3*, *LjFEN1*, *LjLAN*, *LjLB1/LjLB2/LjLB3*, *LjNAP1*, *LjPIR*, *LjSCARN*, *LjSYMRK*, *MtCCaMK/LOF*, *MtDMI1/LjPOLLUX*, *MtNSP1*, *PvAGO5/GmAGO5*, *PvRabA2*, and *LjNUP85* (Supplemental Figure 26, type 3). The copy numbers of homologs in SNF clades for 30 key nodulation genes (Supplemental Figure 26) were estimated from molecular phylogenetic analyses (Supplemental Figures 28–57). Non-nodulating legumes have significantly fewer genes in the SNF clades than nodulating/nitrogen-fixing legumes (*t*-test; $p < 0.001$). It is possible that nodulating legumes have gained such genes or that non-nodulating legumes have lost them.

Fabaceae have experienced multiple WGDs, including those at the origins of the two largest subfamilies and even at the MRCA of Fabaceae (Cannon et al., 2015; Koenen et al., 2020b) (Figures 1, 2, 3, 4, and 5), raising the possibility that some of the WGDs may have increased the copy numbers of nodulation/nitrogen fixation-related genes, thereby facilitating functional evolution. For example, some GDs in early Fabaceae history may have contributed to the switch from actinorhizal to rhizobial nodulation. Indeed, several SNF genes were duplicated in early Fabaceae evolution, including at the MRCA of Fabaceae (*LjNFR1/MtLYK3*, *MtLYK1/MtLYK9*, *LjLNP*, *LjAPN1*, and *LjFEN1* in the SNF clade of Supplemental Figure 26), supporting SH1 in which the switch to rhizobial nodulation occurred at the MRCA

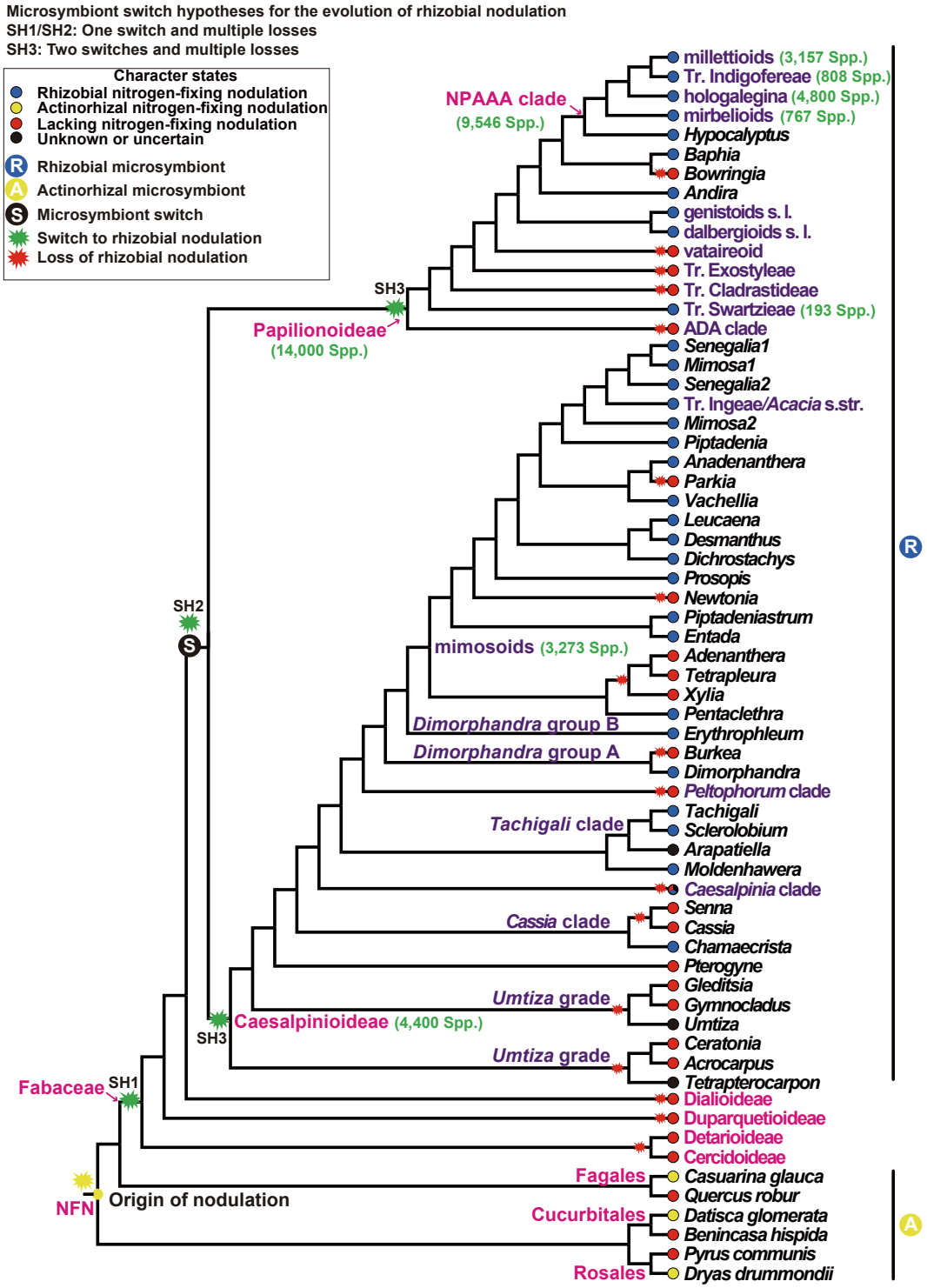

Figure 7. An overview of proposed evolutionary histories of the rhizobial nitrogen-fixing symbiosis in Fabaceae.
This summary tree is a simplified version with tips collapsed as noted in the tip names.
Three slightly different hypotheses for switch(es) from actinorhizal to rhizobial nodulation with subsequent multiple losses are marked as SH1, SH2, and SH3 next to the affected nodes. NFN indicates the nitrogen-fixing clade. The origin of rhizobial nodulation for the nitrogen-fixing clade and multiple losses of rhizobial nodulation are marked with green and red asterisks, respectively.

of Fabaceae or even earlier (Figure 7). The *LjSEN1* gene was duplicated at the MRCA of Caesalpinioideae and Papilionoideae, providing support for SH2. Other genes were duplicated at the MRCA of Caesalpinioideae and/or at the MRCA of Papilionoideae (Supplemental Figure 26, type 4), supporting SH3 (Figure 7) and the previous proposal that the

ancestral Papilionoideae WGD contributed to an increased capacity to form nitrogen-fixing nodules (Young et al., 2011; Li et al., 2013).

Implications for legume diversity

Phylotranscriptomic analyses with 463 legumes resulted in a well-resolved Fabaceae phylogeny for subfamilies, tribes and other multi-generic clades, and many genera. The phylogenetic relationships in subdivisions of the two largest subfamilies often take the form of a grade of relatively species-poor lineages that are successive sisters to one of the large legume clades. For example, the tribe-level subgroups in Caesalpinioideae form a grade of nine lineages outside the mimosoids, and within the mimisoids, 14 lineages form a grade outside the clade comprising most of Ingeae and *Acacia* s.s. (Figures 1 and 2). Also, five lineages with relatively few species of Papilionoideae form a grade outside the clade that contains the rest of this subfamily (Figures 1, 3, 4, and 5). The pattern of a grade of species-poor groups as successive sisters to a much larger clade suggests that degrees of diversification of two sister clades are dramatically different following each of the many divergences in the two largest Fabaceae subfamilies. Nonetheless, possible biases in sequences for phylogenetic analyses cannot be ruled out.

A possible factor in Fabaceae diversification is the environment in which the early Fabaceae diversified. Our analyses suggest that Fabaceae originated within ~4 mya before the K/Pg boundary, with rapid divergences of subfamilies near the K/Pg boundary, suggesting that the early Fabaceae survived the environmental changes and benefited from the niches available after mass extinctions. Further diversifications among most tribes and other multi-generic clades were estimated to have occurred from 65 to 50 mya, coinciding with overall global warming during the Paleocene-Eocene Thermal Maximum and the Early Eocene Climatic Optimum (Figure 6). The role of increased temperature as a driver of biodiversity was previously acknowledged (Erwin, 2009; Benton, 2010; Condamine et al., 2013) and is thought to reflect increased rates of biological processes, shortened generation times, and the consequent higher speciation rates under higher temperatures.

However, environmental changes do not easily explain the differences in species diversity (number) among Fabaceae subfamilies (from a single species in Duparquetioideae to ~14 000 species in Papilionoideae) or between sister lineages with different species richness, especially among the grades of Caesalpinioideae and Papilionoideae. It is possible that WGDs associated with episodes of rapid global change may have promoted legume diversification (Koenen et al., 2020b). Here, we report clusters of large numbers of GDs that support numerous WGDs in legumes. We offer support from more taxa and more precise phylogenetic placement of previously proposed WGDs at the MRCAs of Caesalpinioideae and Papilionoideae, and we propose new WGDs for Fabaceae and other subgroups. Among the 28 WGD/WGT events detected here, 17 (59%) and 6 (21%) are in Papilionoideae and Caesalpinioideae, respectively, suggesting that genes from WGD events may have contributed to increased diversification in these Fabaceae lineages. In particular, the retained duplicates from WGDs may have enhanced the ability of early legumes to survive stressful environments and to evolve new functions; this idea is supported by the GO category enrichment analysis. Specifically, as also suggested in previous studies (Young et al., 2011; Li et al., 2013), the ancestral polyploidy event may have led to enhanced root nodule symbiosis in the Papilionoideae by increasing the copy number of nodulation genes.

Moreover, a comparison between estimated divergence times and detected WGD events places six events (WGD1–6) near the K/Pg boundary (~65 mya), one (WGD7) in the Paleocene (65–55 mya), eight (WGD8–10, 12, 13, 15, 16, 20) during the Eocene (55–33 mya), nine (WGD11, 14, 17, 18, 21, 23, 25, 27, 28) during the Oligocene (33–23 mya), and five (WGD19, 22, 24, 26) in more recent times (23–0 mya). The six WGDs close to the K/Pg boundary occurred at the ancestors of Fabaceae or one of the subfamilies and may have helped legumes survive this period of great stress, as previously proposed for other WGDs (Vanneste et al., 2014). However, a possible role for polyploidy in the survival of legumes beyond the K/Pg boundary is contingent upon whether the WGD took place prior to, concomitant with, or after the global events surrounding the K/Pg boundary, as the dating of polyploidy events is not a trivial task (Doyle and Egan, 2010). Although polyploidy in legumes prior to the K/Pg boundary may have helped them survive that event, it is also possible that the severe environmental changes and stresses that caused mass extinctions at the K/Pg boundary also triggered widespread genomic changes, including legume polyploidy. This idea is supported by the finding that stressful environments can increase the likelihood of unreduced gamete formation, thereby increasing polyploidy (Mason and Pires, 2015), and also by a non-random spike in polyploid formations during this period (Lohaus and Van de Peer, 2016).

The nine WGDs during the Paleocene and Eocene may have contributed to the adaptation of legumes to tropical and subtropical climates, as the relevant legumes would have to survive through periods of global warming (e.g., the Paleocene-Eocene Thermal Maximum and the Early Eocene Climatic Optimum). Subsequently, another nine WGDs occurred in the Oligocene during the global expansion of grassland, increasing aridity, and seasonal climate changes. Most of the sampled Fabaceae genera diverged during this period, suggesting that the increased genetic materials from WGDs may have promoted plant morphological changes. The combination of enhanced survival through stressful periods and adaptive morphological changes (such as transitions to herbs and zygomorphic flowers) may have contributed to the great diversity in Fabaceae.

One important trait shared by most (but not all) Fabaceae species is rhizobial nodulation for nitrogen fixation, which accounts for 92.27% of Fabaceae species diversity. The adaptability for growth in nitrogen-poor soils afforded by rhizobial nodulation (Spehn et al., 2002) may have contributed to species diversity (Vandermeer, 1989; Vandermeer et al., 1990), as supported by the much smaller sizes of the four subfamilies that lack rhizobial nodulation compared with the two subfamilies with species that can form rhizobial nodules. Furthermore, within Caesalpinioideae and Papilionoideae, early divergent tribe-level

subgroups that lack nodulation are also relatively small, further supporting the idea that nodulation-enabled nitrogen fixation contributed to species diversity. The estimated ages of the two clades (one each in Caesalpinioideae and Papilionoideae) with the most nodulating species were ~60 mya (Figure 6) when many tribe-level subgroups diverged rapidly, supporting nitrogen fixation as a factor that promotes species diversification and richness. Recently, an examination of over 500 legume genera with nodulation information found a link between nodulation and greater species richness, but not necessarily with net diversification rate (Afkhami et al., 2018).

The ancestral character reconstruction here suggests multiple origins of rhizobial symbiotic nodules for nitrogen fixation in the two largest Fabaceae subfamilies, as previously hypothesized (Doyle, 1994, 1998, 2011). However, this hypothesis does not consider the fact that rhizobial nodulation (or nodulation in general) is a complex process that requires many genes, making the scenario of multiple independent nodulation origins very unlikely (van Velzen et al., 2019). An alternative hypothesis of a single nodulation origin followed by massive, multiple losses has received renewed attention and increased acceptance. It has support from genome-wide comparative analyses (Doyle, 2016; Griesmann et al., 2018; van Velzen et al., 2019), which revealed that the *NIN* and *RPG* genes for nitrogen-fixing nodulation were lost independently in non-nodulating relatives of nitrogen-fixing species. As further support, our gene family analysis of key nodulation genes from 19 legumes confirmed the loss of *NIN* and *RPG* (Griesmann et al., 2018) and provided the first evidence for the loss of *NFR1*, another crucial nodulation gene.

Among the four orders of the nitrogen-fixing clade, Fabales (with rhizobial nodulating legumes) are sister to Fagales, whereas Rosales are sister to Cucurbitales; furthermore, most non-legume nitrogen fixers form actinorhizal nodules, suggesting that actinorhizal nodulation is ancestral. Thus, there would be one or a few switch(es) from actinorhizal to rhizobial nodulation after the divergence of Fabales from other orders in the nitrogen-fixing clade and before (or shortly after) the separation of Caesalpinioideae and Papilionoideae. Our sequence comparisons and phylogenetic analyses of 30 gene families with known nodulation-related genes found shared internal in-frame deletions in the legume homologs of *NIN* and *RPG* genes and GDs that support the SH1 hypothesis in which the switch to rhizobial nodulation occurred at the Fabaceae ancestor. GDs in other genes support other hypotheses for switches either at the MRCA of the combined Caesalpinioideae + Papilionoideae clade (SH2) or separately at the ancestors of Caesalpinioideae and Papilionoideae (SH3). These switch(es) to rhizobial nodulation were all followed by multiple losses of nodulation (and nitrogen fixation), suggesting that the ancestors of multiple lineages may have lived in environments with relatively abundant fixed nitrogen. It has been proposed that changes in atmospheric CO_2 levels have driven the evolution of plant anatomy and physiology (van Velzen et al., 2019). Nitrogen-fixing nodulation has high carbon and energy costs and would only be advantageous when plant growth was limited by nitrogen and when the benefit of SNF outweighed the carbon costs of symbiotic rhizobia. Conversely, when CO_2 levels and photosynthesis decreased and the carbon cost of nodulation outweighed its benefits, nodulation would be inactivated, initially reversibly. In cases of persistent inactivation of nodulation, the inactive genes can accumulate mutations, resulting in a loss of the ability to nodulate (van Velzen et al., 2019). Therefore, the decreases in CO_2 levels starting at ~47 mya and again at ~34 mya (Beerling and Royer, 2011) after the diversification of the Caesalpinioideae and Papilionoideae lineages (Figure 6) may have been an important environmental factor that contributed to multiple losses of rhizobial nodulation. Subsequently, taxa capable of nitrogen fixation (Figure 7; e.g., the NPAAA clade and most of the mimosoids) further diversified much more extensively than the non-fixers, suggesting that nitrogen fixation by rhizobial nodulation may have contributed to the subsequent adaptation to a greater number of niches, as discussed previously (Afkhami et al., 2018).

In short, this study has established a well-resolved Fabaceae phylogeny, providing a valuable foundation for many evolutionary and comparative studies. Analyses of divergence times, WGDs, and the evolutionary history of the nodulation symbiosis revealed striking coincidences that may support the idea that environmental and genomic/genic changes combined to promote the diversification of legume lineages. Large amounts of legume sequence data and phylogenomic analyses provide the phylogenetic and molecular contexts for understanding the evolution and ecology of one of the most important symbiotic processes between plants and bacteria.

METHODS

Taxon sampling, sequencing, and transcriptome and genome assembly

The species sampled here included legume members and taxa in Fabales, Fagales, and other eudicot orders (Supplemental Table 1). A total of 463 legumes were included (Supplemental Table 1), representing all 6 subfamilies and covering 59 tribes or rank-free clades, 333 genera, and 12 unassigned species. In addition, 16 other eudicot species were used as outgroups, including 4 Fabales (Polygalaceae, Quillajaceae), 4 Fagales, and 8 species from 3 other orders.

Young leaves, buds, or seedlings were collected for DNA/RNA extraction. Transcriptomes and genomes were *de novo* assembled into contigs using Trinity v2.9.0 (Grabherr et al., 2011), SOAPdenovo2 (Luo et al., 2012), and TGICLv2.1 (Pertea et al., 2003) with the parameters described previously (Huang et al., 2016a). TransDecoder (http://transdecoder.sourceforge.net/) was used to predict CDS regions, and redundant contigs from each sample were reduced using CD-HIT 4.6 (Fu et al., 2012) with the parameter -c 0.98 as in previous studies (Huang et al., 2016a; Xiang et al., 2017; Zeng et al., 2017; Qi et al., 2018).

Identification of low-copy candidate orthologs

The reservoir of putative low-copy nuclear genes used in this study was integrated from three separate gene sets (A, B, and C) (Supplemental Figure 2A), a strategy that has been useful in legumes (Vatanparast et al., 2018). Redundant OGs with the same gene ID corresponding to model plants (*Arabidopsis thaliana* and *Glycine max*) were removed (Supplemental Table 4). The resulting 2833 OGs were used as seed genes to obtain the corresponding putative orthologs (E-value < $1e^{-20}$) from 484 samples in HaMStR v13.2.3 (Ebersberger et al., 2009). Subsequently, 1559 OGs (set 1) were selected (Supplemental Figure 2), aligned using MAFFT with default settings (Katoh and Standley, 2013), and trimmed using trimAl v1.2 with default settings (Capella-Gutierrez et al., 2009). Next, additional filtering based on taxon coverage, alignment length, and other parameters yielded 6 smaller sets (sets 2 to 7) of 1083 to 131 OGs (Supplemental Figure 2 and Supplemental Table 3 and see the Supplemental text for details). The removal of

sequences with relatively low taxon coverage and short alignment regions to obtain successively smaller gene sets is effective in reducing noise and errors and facilitates the reconstruction of a robust phylogeny from coalescent analyses (Supplemental Table 12).

Phylogenetic analysis

We obtained coalescent trees for gene sets 1 through 7 and summarized the topologies from the seven coalescent trees to propose a final model tree (Figures 1, 2, 3, 4, and 5 and Supplemental Figure 3). Amino acid sequences were aligned using MAFFT v7 (Katoh and Standley, 2013) with the "–auto" parameter. Poorly aligned regions were further trimmed using trimAl v1.2 software (Capella-Gutierrez et al., 2009) with the "-automated1″ parameter. Multiple amino acid sequence alignments were converted to nucleotide alignments with PAL2NAL software (Suyama et al., 2006). Single-gene trees were reconstructed with RAxML v8.2.12 (Stamatakis, 2014) under the GTRCAT model. For each gene group, 100 bootstrap replicates were generated for coalescent analysis with ASTRAL-III (Zhang et al., 2018). The 1559 orthologous genes and the 1559 corresponding single-gene ML trees of 483 species are available at a GitHub database (https://github.com/Genomic-docker/Phylogentic-gene-markers-in-Fabaceae-phylogenomics-online.website).

Reconstruction of ancestral nodulation states

All codings for nodulation status are provided in Supplemental Table 9. The ancestral characters were traced using maximum parsimony in Mesquite v3.04 (Maddison and Maddison, 2007) based on the topology in Figure 1. We also reconstructed the ancestral states by marginal probabilities using the rayDISC function (ML) of corHMM in R (Beaulieu et al., 2013). Estimations were performed with all combinations of model and root.p options (Beaulieu et al., 2013). We then selected the best result by evaluating AIC_C scores and weight, and the result shown was reconstructed by estimating the marginal probability of node states according to Yang (2006) under the assumption that the rates between two states do not differ. The root state probability was determined by the procedure described by Maddison et al. (2007) and FitzJohn et al. (2009).

Divergence time estimation

We calibrated a PL-based (Sanderson, 2002; Knapp et al., 2005) molecular clock using 23 recognized fossil-based age constraints (Supplemental Table 5) for dating analyses in treePL (Smith and O'Meara, 2012). An ML phylogenetic reconstruction was performed using the concatenation of OG set 7 (131 genes, 181 544 sites) and the topology of the summarized phylogeny (Figure 1), generating a tree with branch lengths for age estimation.

Phylogenomics for detection of GD clusters and other analyses for WGD

The legume species were divided into 12 overlapping groups and 1 overall group with representatives of major lineages (Supplemental Figures 8–16 and Supplemental Figures 18–21) based on our phylogeny. Gene family sequences were identified by all-against-all BLASTP searches followed by clustering with the MCL algorithm (Enright et al., 2002) (https://micans.org/mcl/) with an inflation value of 6.0. Each cluster of at least five taxa was aligned using MAFFT (https://mafft.cbrc.jp) with the "–auto" parameter and used to construct ML trees with the maximum likelihood method IQ-TREE (Nguyen et al., 2015) (http://www.iqtree.org/). BS values were estimated from 1000 replicates using the GTRGAMMA model. The gene trees were then mapped with the species tree to identify duplicates using the phylogenomic tool tree2gd v2.4 (custom software available from https://sourceforge.net/projects/tree2gd/ or https://github.com/Dee-chen/Tree2gd). Only those duplicates that were shared by at least two species and had >70% bootstrap support at the node of the paralogs were counted. Furthermore, the number of duplicated species in each of two paralogous subclades was required to be more than 20% of the total species in the group. Nodes associated with GD numbers >450 were retained for further evaluation by considering the GD types. Proposed WGDs that were shared by species with sequenced genome(s) were also evaluated by collinearity (synteny) analyses in MCScanX (Wang et al., 2012) (http://chibba.pgml.uga.edu/mcscan2) with default settings. The synonymous mutation ratio (Ks) was calculated for syntenic gene pairs derived from plant genomes using KaKs_calculator v2.0 with the NG method (model of Nei-Gojobori) (Wang et al., 2010). GO analysis was performed using agriGO on the retained duplicates derived from the MRCA of Fabaceae (Du et al., 2010) (http://bioinfo.cau.edu.cn/agriGO/).

Analyses of polyploidization event(s) in early Fabaceae history

Multiple analyses were performed to test the mode of the polyploidy event at the MRCA of Fabaceae. First, 5088 gene family trees with putative GDs (without duplicated species coverage filtration) were obtained from the phylogenetic analyses of genes from 70 species (including 65 legumes) (Supplemental Figure 21) and examined for GDs that formed a cluster at the MRCA of Fabaceae. Gene families with GDs were examined for duplication and taxon patterns (Supplemental Figure 58; sps.duplication, sps.sister, or sps.other). For each species, the proportions of gene families that belonged to each of the three types were determined (Supplemental Figure 58).

Second, gene families with GDs mapped to the MRCA of Fabaceae were examined to count the numbers of specific gene families with topologies/subfamily distributions that supported alternative hypotheses for polyploidy in early legume history. To accommodate the possible impact of failure in gene detection due to incomplete genome/transcriptome sequencing and incorrect assembly or annotation, we examined four gene sets (gene set1–4) with different numbers of legume representatives and different filtering criteria for GD detection. Gene set1 was obtained from the gene family trees of 65 legumes (Supplemental Figure 21, those mentioned in the first analysis above) using criteria of BS of node/subclades (a duplicated node and its two subclades) $\geq$50 and number of species with duplicates $\geq$4. Gene set2 was based on the criteria of BS of node $\geq$70 and the number of species with duplicates $\geq$5. Gene set3 was inferred from an analysis of genes from 16 species with BS of node/subclades $\geq$50 and species with duplicates $\geq$2 (Supplemental Figure 60A). Gene set4 was inferred from an analysis of 48 species with BS of node/subclades $\geq$50 and species with duplicates $\geq$2 (Supplemental Figure 61). The selected legumes had relatively high unigene N50 values or BUSCO (Simão et al., 2015) completeness assessments (quality information for the assemblies is provided in Supplemental Table 2). Example topologies that support allopolyploidy for a subset of Fabaceae subfamilies or polyploidy at the MRCA of Fabaceae are illustrated in Supplemental Figure 59, and 140 possible gene family distributions that support WGD1 at the stem of Fabaceae are listed in Supplemental Table 13.

Finally, the tree reconciliation-based method GRAMPA (Gregg et al., 2017) was implemented to detect evidence for possible allopolyploidy. GRAMPA tests hypotheses of polyploidy by reconciling singly labeled gene trees to different multi-labeled trees. The multi-labeled tree with the lowest score indicates the polyploid species and their closest surviving parental lineages (Thomas et al., 2017). The analysis was performed using gene set3 (Supplemental Figure 60, described above) with sequences from 5 genomes and 11 transcriptomes that had relatively high unigene N50 values or BUSCO completeness assessments (2 Papilionoideae, 2 Caesalpinioideae, 3 Dialioideae, 3 Cercidoideae, 4 Detarioideae, and 2 outgroups; information on the assemblies is provided in Supplemental Table 2).

Molecular evolution of key nodulation genes

Twenty-eight publicly available genomes and 5 publicly available transcriptomes with high BUSCO percentages were used for molecular evolution analyses; these included 2 non-nodulating legumes, *Cercis canadensis* (97.7%) and *Nissolia schottii* (94.5%) (Supplemental Table 10). Nodulation-related genes were identified by all-against-all BLASTP

searches with an E-value less than $1e^{-5}$ and greater than 20% amino acid sequence identity. Sequences were aligned by MAFFT (Katoh and Standley, 2013) with accurate aligning options "–maxiterate 1000 –localpair," manually adjusted with AliView (Larsson, 2014), and trimmed with trimAl v1.4 (Capella-Gutierrez et al., 2009) using the "-gt 0.1" option. An alignment of CDS nucleotide sequences was obtained from the amino acid sequence alignment using PAL2NAL v13 (Suyama et al., 2006). ModelFinder (http://www.iqtree.org/ModelFinder/) was used to select the best model under the Bayesian information criterion. IQ-TREE (http://www.iqtree.org/) was then used to reconstruct all ML phylogenetic trees with the model suggested by Modelfinder and to perform a bootstrap significance test with 1000 bootstrap replicates.

See the Supplemental text for additional information on methods.

SUPPLEMENTAL INFORMATION

Supplemental information is available at *Molecular Plant Online*.

FUNDING

This work was supported by funds from the National Natural Science Foundation of China (31770242 and 31970224), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB31010000), funds from the State Key Laboratory of Genetic Engineering and the Ministry of Education Key Laboratory of Biodiversity Science and Ecological Engineering at Fudan University, and funds from the Pennsylvania State University. We also gratefully acknowledge the China Scholarship Council for financial support to Y.Z. for collaboration at the Pennsylvania State University.

AUTHOR CONTRIBUTIONS

H.M., T.-S.Y., and C.-H.H. designed the study and managed the project. A.N.E. aided in taxon sampling design. H.M., Y.Z., T.-S.Y., R.Z., Y.H., and J.G. collected materials and performed species identification. Y.Z., R.Z., and Y.H. isolated RNA for some taxa. Y.Z. performed raw data analysis, transcriptome assembly, gene prediction, gene annotation, phylogenetic analysis, molecular clock estimation, reconstruction of ancestral states of rhizobial nodulation status, detection of WGDs, and nodulation gene family analyses with help from J.Q. and C.-H.H. T.Z., R.Z., and K.J. assisted in figure preparation. R.Z., T.-S.Y., and Y.Z. performed classification. Y.Z. wrote the manuscript. C.-H.H., H.M., T.-S.Y., and A.N.E. revised the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We thank Dr. Weibin Xu (Guangxi Institute of Botany, Chinese Academy of Sciences), Dr. Xinxin Zhu (Xinyang Normal University), Dr. Yingxiang Wang (Fudan University), and Dr. Pablo Bolaños-Villegas and Mr. Sergio Castro-Pacheco (University of Costa Rico) for assistance in taxon sampling. In addition, we thank Dr. Manuel de la Estrella for kindly providing the flower photograph of *Duparquetia* in Figure 1. We are particularly thankful for the valuable comments on the manuscript from three anonymous reviewers. No conflict of interest declared.

Received: March 6, 2020
Revised: July 31, 2020
Accepted: February 19, 2021
Published: February 22, 2021

REFERENCES

Afkhami, M.E., Mahler, D.L., Burns, J.H., Weber, M.G., Wojciechowski, M.F., Sprent, J., and Strauss, S.Y. (2018). Symbioses with nitrogen-fixing bacteria: nodulation and phylogenetic data across legume genera. Ecology **99**:502.

Barker, D., Bianchi, S., Blondon, F., Dattée, Y., Duc, G., Essad, S., Flament, P., Gallusci, P., Génier, G., Guy, P., et al. (1990). *Medicago truncatula*, a model plant for studying the molecular genetics of the *Rhizobium*-legume symbiosis. Plant Mol. Biol. Rep. **8**:40–49.

Battenberg, K., Potter, D., Tabuloc, C.A., Chiu, J.C., and Berry, A.M. (2018). Comparative transcriptomic analysis of two actinorhizal plants and the legume *Medicago truncatula* supports the homology of root nodule symbioses and is congruent with a two-step process of evolution in the nitrogen-fixing clade of angiosperms. Front. Plant Sci. **9**:1256.

Beaulieu, J.M., O'Meara, B.C., and Donoghue, M.J. (2013). Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in Campanulid angiosperms. Syst. Biol. **62**:725–737.

Beerling, D.J., and Royer, D.L. (2011). Convergent cenozoic CO_2 history. Nat. Geosci. **4**:418–420.

Bell, E.A. (1981). Non-protein amino acids in the Leguminosae. In Advances in Legume Systematics, R.M. Polhill and P.H. Raven, eds. (Kew, UK: Royal Botanic Gardens), pp. 489–499, part 2.

Benton, M.J. (2010). The origins of modern biodiversity on land. Philos. Trans. R. Soc. Lond. Ser. B: Biol. Sci. **365**:3667–3679.

Bertioli, D.J., Cannon, S.B., Froenicke, L., Huang, G., Farmer, A.D., Cannon, E.K.S., Liu, X., Gao, D., Clevenger, J., Dash, S., et al. (2016). The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. Nat. Genet. **48**:438–446.

Birky, C.W. (2001). The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. Annu. Rev. Genet. **35**:125–148.

Bouchenak-Khelladi, Y., Maurin, O., Hurter, J., and Van der Bank, M. (2010). The evolutionary history and biogeography of Mimosoideae (Leguminosae): an emphasis on frican acacias. Mol. Phylogen. Evol. **57**:495–508.

Bozsoki, Z., Gysel, K., Hansen, S.B., Lironi, D., Krönauer, C., Feng, F., de Jong, N., Vinther, M., Kamble, M., Thygesen, M.B., et al. (2020). Ligand-recognizing motifs in plant LysM receptors are major determinants of specificity. Science **369**:663–670.

Bruneau, A., Forest, F., Herendeen, P.S., Klitgaard, B.B., and Lewis, G.P. (2001). Phylogenetic relationships in the Caesalpinioideae (Leguminosae) as inferred from chloroplast *trnL* intron sequences. Syst. Bot. **26**:487–515.

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

Buendia, L., Girardin, A., Wang, T., Cottret, L., and Lefebvre, B. (2018). LysM receptor-like kinase and LysM receptor-like protein families: an update on phylogeny and functional characterization. Front. Plant Sci. **9**:1531.

Cannon, S.B., McKain, M.R., Harkess, A., Nelson, M.N., Dash, S., Deyholos, M.K., Peng, Y., Joyce, B., Stewart, C.N., Rolf, M., et al. (2015). Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. Mol. Biol. Evol. **32**:193–210.

Cannon, S.B., Sterck, L., Rombauts, S., Sato, S., Cheung, F., Gouzy, J., Wang, X., Mudge, J., Vasdewani, J., Schiex, T., et al. (2006). Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. Proc. Natl. Acad. Sci. U S A **103**:14959–14964.

Capella-Gutierrez, S., Silla-Martinez, J.M., and Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics **25**:1972–1973.

Cardoso, D., De Queiroz, L.P., Pennington, R.T., De Lima, H.C., Fonty, É., Wojciechowski, M.F., and 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., and Lavin, M. (2013). Reconstructing the deep-branching relationships of the papilionoid legumes. S. Afr. J. Bot. **89**:58–75.

Chappill, J.A., and Maslin, B.R. (1995). A phylogenetic assessment of tribe Acacieae. In Advances in Legume Systematics, M.D. Crisp and J.J. Doyle, eds. (Kew, UK: Royal Botanic Gardens), pp. 77–99, part 7.

Clavijo, F., Diedhiou, I., Vaissayre, V., Brottier, L., Acolatse, J., Moukouanga, D., Crabos, A., Auguy, F., Franche, C., Gherbi, H., et al. (2015). The *Casuarina NIN* gene is transcriptionally activated throughout *Frankia* root infection as well as in response to bacterial diffusible signals. New Phytol. **208**:887–903.

Condamine, F.L., Rolland, J., and Morlon, H. (2013). Macroevolutionary perspectives to environmental change. Ecol. Lett. **16**:72–85.

Davis, C.C., Xi, Z., and Mathews, S. (2014). Plastid phylogenomics and green plant phylogeny: almost full circle but not quite there. BMC Biol. **12**:11.

de la Estrella, M., Forest, F., Klitgard, B., Lewis, G.P., Mackinder, B.A., de Queiroz, L.P., Wieringa, J.J., and 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.

De Mita, S., Streng, A., Bisseling, T., and Geurts, R. (2014). Evolution of a symbiotic receptor through gene duplications in the legume–rhizobium mutualism. New Phytol. **201**:961–972.

de Queiroz, L.P., Pastore, J.F., Cardoso, D., Snak, C., de, C.L.A.L., Gagnon, E., Vatanparast, M., Holland, A.E., and Egan, A.N. (2015). A multilocus phylogenetic analysis reveals the monophyly of a recircumscribed papilionoid legume tribe Diocleae with well-supported generic relationships. Mol. Phylogen. Evol. **90**:1–19.

Dos Santos, P.C., Fang, Z., Mason, S.W., Setubal, J.C., and Dixon, R. (2012). Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. BMC Genomics **13**:162.

Doyle, J., Biens, P., Doerenkamp, A., and Jardiné, S. (1977). Angiosperm pollen from the pre-Albian Lower Cretaceous of Equatorial Africa. Bull. Cent. Rech. Explor. Prod. Elf-aquitaine **1**:451–473.

Doyle, J.J. (1994). Phylogeny of the legume family: an approach to understanding the origins of nodulation. Annu. Rev. Ecol. Syst. **25**:325–349.

Doyle, J.J. (1998). Phylogenetic perspectives on nodulation: evolving views of plants and symbiotic bacteria. Trends Plant Sci. **3**:473–478.

Doyle, J.J. (2011). Phylogenetic perspectives on the origins of nodulation. Mol. Plant Microbe Interact. **24**:1289–1295.

Doyle, J.J. (2012). Polyploidy in legumes. In Polyploidy and Genome Evolution–Soltis, P.S.. and D.E. Soltis, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 147–180.

Doyle, J.J. (2016). Chasing unicorns: nodulation origins and the paradox of novelty. Am. J. Bot. **103**:1865–1868.

Doyle, J.J., and Egan, A.N. (2010). Dating the origins of polyploidy events. New Phytol. **186**:73–85.

Du, Z., Zhou, X., Ling, Y., Zhang, Z., and Su, Z. (2010). agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res **38**:W64–W70.

Ebersberger, I., Strauss, S., and von Haeseler, A. (2009). HaMStR: profile hidden Markov model based search for orthologs in ESTs. BMC Evol. Biol. **9**:157.

Egan, A.N., and Doyle, J. (2010). A comparison of global, gene-specific, soybean (*Glycine max*). Syst. Biol. **59**:534–547.

Egan, A.N., and Vatanparast, M. (2019). Advances in legume research in the genomics era. Aust. Syst. Bot. **32**:459–483.

Egan, A.N., Vatanparast, M., and Cagle, W. (2016). Parsing polyphyletic *Pueraria*: delimiting distinct evolutionary lineages through phylogeny. Mol. Phylogen. Evol. **104**:44–59.

Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res **30**:1575–1584.

Erwin, D.H. (2009). Climate as a driver of evolutionary change. Curr. Biol. **19**:R575–R583.

FitzJohn, R.G., Maddison, W.P., and Otto, S.P. (2009). Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Syst. Biol. **58**:595–611.

Freeling, M., and Thomas, B.C. (2006). Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res. **16**:805–814.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics **28**:3150–3152.

Garg, N., and Geetanjali. (2009). Symbiotic nitrogen fixation in legume nodules: process and signaling: a review. In Sustainable Agriculture, E. Lichtfouse, M. Navarrete, P. Debaeke, S. Veronique, and C. Alberola, eds. (Netherlands: Springer), pp. 519–531.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol. **29**:644–652.

Gregg, W.T., Ather, S.H., and Hahn, M.W. (2017). Gene-tree reconciliation with MUL-trees to resolve polyploidy events. Syst. Biol. **66**:1007–1018.

Griesmann, M., Chang, Y., Liu, X., Song, Y., Haberer, G., Crook, M.B., Billault-Penneteau, B., Lauressergues, D., Keller, J., Imanishi, L., et al. (2018). Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. Science **361**:eaat1743.

Guo, J., Xu, W., Hu, Y., Huang, J., Zhao, Y., Zhang, L., Huang, C.-H., and Ma, H. (2020). Phylotranscriptomics in Cucurbitaceae reveal multiple whole-genome duplications and key morphological and molecular innovations. Mol. Plant **13**:1–17.

Handberg, K., and Stougaard, J. (1992). *Lotus japonicus*, an autogamous, diploid legume species for classical and molecular genetics. Plant J. **2**:487–496.

Hane, J.K., Ming, Y., Kamphuis, L.G., Nelson, M.N., Garg, G., Atkins, C.A., Bayer, P.E., Bravo, A., Bringans, S., Cannon, S., et al. (2017). A comprehensive draft genome sequence for lupin (*Lupinus angustifolius*), an emerging health food: insights into plant-microbe interactions and legume evolution. Plant Biotechnol. J. **15**:318–330.

Herendeen, P.S., Bruneae, A., and Lewis, G.P. (2003). Phylogenetic relationships in caesalpinioid legumes: a preliminary analysis based on morphological and molecular data. In Advances in Legume Systematics, Higher Level Systematics, B.B. Klitgaard and A. Brunneau, eds. (Kew, UK: Royal Botanic Gardens), pp. 37–62, part 10.

Hu, J.M., Lavin, M., Wojciechowski, M.F., and Sanderson, M.J. (2000). Phylogenetic systematics of the tribe Millettieae (Leguminosae) based on chloroplast *trnK/matK* sequences and its implications for evolutionary patterns in Papilionoideae. Am. J. Bot. **87**:418–430.

Huang, C.H., Sun, R., Hu, Y., Zeng, L., Zhang, N., Cai, L., Zhang, Q., Koch, M.A., Al-Shehbaz, I., Edger, P.P., et al. (2016a). Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. Mol. Biol. Evol. **33**:394–412.

Huang, C.H., Zhang, C., Liu, M., Hu, Y., Gao, T., Qi, J., and Ma, H. (2016b). Multiple polyploidization events across Asteraceae with two nested events in the early history revealed by nuclear phylogenomics. Mol. Biol. Evol. **33**:2820–2835.

Hughes, C., and Eastwood, R. (2006). Island radiation on a continental scale: exceptional rates of plant diversification after uplift of the Andes. Proc. Natl. Acad. Sci. U S A **103**:10334–10339.

Jiao, Y., Leebens-Mack, J., Ayyampalayam, S., Bowers, J.E., McKain, M.R., McNeal, J., Rolf, M., Ruzicka, D.R., Wafula, E., Wickett, N.J., et al. (2012). A genome triplication associated with early diversification of the core eudicots. Genome Biol. **13**:R3.

Jiao, Y., Li, J., Tang, H., and Paterson, A.H. (2014). Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. Plant Cell **26**:2792–2802.

Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P., Hu, Y., Liang, H., Soltis, P.S., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. Nature **473**:97–100.

Käss, E., and Wink, M. (1996). Molecular evolution of the Leguminosae: phylogeny of the three subfamilies based on *rbcL*-sequences. Biochem. Syst. Ecol. **24**:365–378.

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

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. **30**:772–780.

Knapp, M., Stöckler, K., Havell, D., Delsuc, F., Sebastiani, F., and Lockhart, P.J. (2005). Relaxed molecular clock provides evidence for long-distance dispersal of *Nothofagus* (southern beech). PLoS Biol. **3**:38–43.

Koenen, E.J.M., Kidner, C., de Souza, É.R., Simon, M.F., Iganci, J.R., Nicholls, J.A., Brown, G.K., de Queiroz, L.P., Luckow, M., Lewis, G.P., et al. (2020). Hybrid capture of 964 nuclear genes resolves evolutionary relationships in the mimosoid legumes and reveals the polytomous origins of a large pantropical radiation. Am J Bot. **107**:1710–1735.

Koenen, E.J., Ojeda, D.I., Steeves, R., Migliore, J., Bakker, F.T., Wieringa, J.J., Kidner, C., Hardy, O.J., Pennington, R.T., and Bruneau, A. (2020a). 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.

Koenen, E.J.M., Ojeda, D.I., Bakker, F.T., Wieringa, J.J., Kidner, C., Hardy, O.J., Pennington, R.T., Herendeen, P.S., Bruneau, A., and Hughes, C.E. (2020b). The origin of the legumes is a complex paleopolyploid phylogenomic tangle closely associated with the Cretaceous-Paleogene (K-Pg) mass extinction event. Syst. Biol. https://doi.org/10.1093/sysbio/syaa041.

Konishi, M., and Yanagisawa, S. (2013). *Arabidopsis* NIN-like transcription factors have a central role in nitrate signalling. Nat. Commun. **4**:1617.

Kyalangalilwa, B., Boatwright, J.S., Daru, B.H., Maurin, O., and van der Bank, M. (2013). Phylogenetic position and revised classification of *Acacia s.l.* (Fabaceae: Mimosoideae) in Africa, including new combinations in *Vachellia* and *Senegalia*. Bot. J. Linn. Soc. **172**:500–523.

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics **30**:3276–3278.

Lavin, M., Herendeen, P.S., and Wojciechowski, M.F. (2005). Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary. Syst. Biol. **54**:575–594.

Leebens-Mack, J.H., Barker, M.S., Carpenter, E.J., Deyholos, M.K., Gitzendanner, M.A., Graham, S.W., Grosse, I., Li, Z., Melkonian, M., Mirarab, S., et al. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. Nature **574**:679–685.

Lewis, G.P., Schrire, B.D., Mackinder, B.A., and Lock, M. (2005). Legumes of the World (Kew, UK: Royal Botanic Gardens).

Li, H.L., Wang, W., Mortimer, P.E., Li, R.Q., Li, D.Z., Hyde, K.D., Xu, J.C., Soltis, D.E., and Chen, Z.D. (2015). Large-scale phylogenetic analyses reveal multiple gains of actinorhizal nitrogen-fixing symbioses in angiosperms associated with climate change. Sci. Rep. **5**:14023.

Li, Q.-G., Zhang, L., Li, C., Dunwell, J.M., and Zhang, Y.-M. (2013). Comparative genomics suggests that an ancestral polyploidy event leads to enhanced root nodule symbiosis in the Papilionoideae. Mol. Biol. Evol. **30**:2602–2611.

Li, Z., De La Torre, A.R., Sterck, L., Cánovas, F.M., Avila, C., Merino, I., Cabezas, J.A., Cervera, M.T., Ingvarsson, P.K., and Van de Peer, Y. (2017). Single-copy genes as molecular markers for phylogenomic studies in seed plants. Genome Biol. Evol. **9**:1130–1147.

Limpens, E., Franken, C., Smit, P., Willemse, J., Bisseling, T., and Geurts, R. (2003). LysM domain receptor kinases regulating rhizobial Nod factor-induced infection. Science **302**:630–633.

Lohaus, R., and Van de Peer, Y. (2016). Of dups and dinos: evolution at the K/Pg boundary. Curr. Opin. Plant Biol. **30**:62–69.

LPWG. (2013). Towards a new classification system for legumes: progress report from the 6th International Legume Conference. S. Afr. J. Bot. **89**:3–9.

LPWG. (2017). A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny. Taxon **66**:44–77.

Lu, L.-M., Mao, L.-F., Yang, T., Ye, J.-F., Liu, B., Li, H.-L., Sun, M., Miller, J.T., Mathews, S., and Hu, H.-H. (2018). Evolutionary history of the angiosperm flora of China. Nature **554**:234–238.

Luckow, M., Miller, J.T., Murphy, D.J., and Livshultz, T. (2003). A phylogenetic analysis of the Mimosoideae (Leguminosae) based on chloroplast DNA sequence data. In Advances in Legume Systematics, Higher Level Systematics, B.B. Klitgaard and A. Bruneau, eds. (Kew, UK: Royal Botanic Gardens), pp. 197–220, part 10.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience **1**:18.

Maddison, W.P., and Maddison, D.R. (2007). Mesquite: a modular system for evolutionary analysis. Version 2.0. http://mesquiteproject.org.

Maddison, W.P., Midford, P.E., and Otto, S.P. (2007). Estimating a binary character's effect on speciation and extinction. Syst. Biol. **56**:701–710.

Mandel, J.R., Dikow, R.B., Siniscalchi, C.M., Thapa, R., Watson, L.E., and Funk, V.A. (2019). A fully resolved backbone phylogeny reveals numerous dispersals and explosive diversifications throughout the history of Asteraceae. Proc. Natl. Acad. Sci. U S A **116**:14083–14088.

Manzanilla, V., and 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.

Mason, A.S., and Pires, J.C. (2015). Unreduced gametes: meiotic mishap or evolutionary mechanism? Trends Genet. **31**:5–10.

Murakami, E., Cheng, J., Gysel, K., Bozsoki, Z., Kawaharada, Y., Hjuler, C.T., Sørensen, K.K., Tao, K., Kelly, S., Venice, F., et al. (2018). Epidermal LysM receptor ensures robust symbiotic signalling in *Lotus japonicus*. ife **7**:e33506.

Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol **32**:268–274.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., et al. (2003). TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics **19**:651–652.

Prenner, G., and Klitgaard, B.B. (2008). Towards unlocking the deep nodes of Leguminosae: floral development and morphology of the enigmatic *Duparquetia orchidacea* (Leguminosae, Caesalpinioideae). Am. J. Bot. **95**:1349–1365.

Qi, X., Kuo, L.Y., Guo, C., Li, H., Li, Z., Qi, J., Wang, L., Hu, Y., Xiang, J., Zhang, C., et al. (2018). A well-resolved fern nuclear phylogeny reveals the evolution history of numerous transcription factor families. Mol. Phylogen. Evol. **127**:961–977.

Ren, R., Wang, H., Guo, C., Zhang, N., Zeng, L., Chen, Y., Ma, H., and Qi, J. (2018). Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. Mol. Plant **11**:414–428.

Roy, S., Liu, W., Nandety, R.S., Crook, A., Mysore, K.S., Pislariu, C.I., Frugoli, J., Dickstein, R., and Udvardi, M.K. (2020). Celebrating 20 years of genetic discoveries in legume nodulation and symbiotic nitrogen fixation. Plant Cell **32**:15–41.

Rutten, L., Miyata, K., Roswanjaya, Y.P., Huisman, R., Bu, F., Hartog, M., Linders, S., van Velzen, R., van Zeijl, A., Bisseling, T., et al. (2020). Duplication of symbiotic lysin motif receptors predates the evolution of nitrogen-fixing nodule symbiosis. Plant Physiol. **184**:1004–1023.

Sanderson, M.J. (2002). Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol. Biol. Evol. **19**:101–109.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., et al. (2010). Genome sequence of the palaeopolyploid soybean. Nature **463**:178–183.

Schmutz, J., McClean, P.E., Mamidi, S., Wu, G.A., Cannon, S.B., Grimwood, J., Jenkins, J., Shu, S., Song, Q., Chavarro, C., et al. (2014). A reference genome for common bean and genome-wide analysis of dual domestications. Nat. Genet. **46**:707–713.

Shen, H., Jin, D., Shu, J.-P., Zhou, X.-L., Lei, M., Wei, R., Shang, H., Wei, H.-J., Zhang, R., Liu, L., et al. (2017). Large-scale phylogenomic analysis resolves a backbone phylogeny in ferns. GigaScience **7**:gix116.

Silveira, F.A., Negreiros, D., Barbosa, N.P., Buisson, E., Carmo, F.F., Carstensen, D.W., Conceição, A.A., Cornelissen, T.G., Echternacht, L., and Fernandes, G.W. (2016). Ecology and evolution of plant diversity in the endangered campo rupestre: a neglected conservation priority. Plant Soil **403**:129–152.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31**:3210–3212.

Smith, S.A., and O'Meara, B.C. (2012). treePL: divergence time estimation using penalized likelihood for large phylogenies. Bioinformatics **28**:2689–2690.

Soltis, D.E., Soltis, P.S., Morgan, D.R., Swensen, S.M., Mullin, B.C., Dowd, J.M., and Martin, P.G. (1995). Chloroplast gene sequence data suggest a single origin of the predisposition for symbiotic nitrogen fixation in angiosperms. Proc. Natl. Acad. Sci. U S A **92**:2647–2651.

Spehn, E., Scherer-Lorenzen, M., Schmid, B., Hector, A., Caldeira, M., Dimitrakopoulos, P., Finn, J., Jumpponen, A., O'donnovan, G., and Pereira, J. (2002). The role of legumes as a component of biodiversity in a cross-European study of grassland biomass nitrogen. Oikos **98**:205–218.

Sprent, J. (2009). Legume Nodulation: A Global Perspective (Chichester, UK: Wiley-Blackwell).

Springer, M.S., DeBry, R.W., Douady, C., Amrine, H.M., Madsen, O., de Jong, W.W., and Stanhope, M.J. (2001). Mitochondrial versus nuclear gene sequences in deep-level mammalian phylogeny reconstruction. Mol. Biol. Evol. **18**:132–143.

Stai, J.S., Yadav, A., Sinou, C., Bruneau, A., Doyle, J.J., Fernández-Baca, D., and Cannon, S.B. (2019). *Cercis*: a non-polyploid genomic relic within the generally polyploid legume family. Front. Plant Sci. **10**:345.

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

Stefanović, S., Pfeil, B.E., Palmer, J.D., and Doyle, J.J. (2009). Relationships among Phaseoloid legumes based on sequences from eight chloroplast regions. Syst. Bot. **34**:115–128.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. **34**:W609–W612.

Suzuki, W., Konishi, M., and Yanagisawa, S. (2013). The evolutionary events necessary for the emergence of symbiotic nitrogen fixation in legumes may involve a loss of nitrate responsiveness of the *NIN* transcription factor. Plant Signal. Behav. **8**:e25975.

Thomas, G.W.C., Ather, S.H., and Hahn, M.W. (2017). Gene-tree reconciliation with MUL-trees to resolve polyploidy events. Syst. Biol. **66**:1007–1018.

Trenchard, L.J., Harris, P.J.C., Smith, S.J., and Pasiecznik, N.M. (2008). A review of ploidy in the genus *Prosopis* (Leguminosae). Bot. J. Linn. Soc. **156**:425–438.

van Velzen, R., Doyle, J.J., and Geurts, R. (2019). A resurrected scenario: single gain and massive loss of nitrogen-fixing nodulation. Trends Plant Sci. **24**:49–57.

van Velzen, R., Holmer, R., Bu, F., Rutten, L., van Zeijl, A., Liu, W., Santuari, L., Cao, Q., Sharma, T., Shen, D., et al. (2018). Comparative genomics of the nonlegume *Parasponia* reveals insights into evolution of nitrogen-fixing rhizobium symbioses. Proc. Natl. Acad. Sci. U S A **115**:E4700–E4709.

van Velzen, R., Holmer, R., Bu, F., Rutten, L., van Zeijl, A., Liu, W., Santuari, L., Cao, Q., Sharma, T., Shen, D., et al. (2017). Parallel loss of symbiosis genes in relatives of nitrogen-fixing non-legume Parasponia. bioRxiv, 169706.

Vandermeer, J.H. (1989). The Ecology of Intercropping (Cambridge: Cambridge University Press).

Vandermeer, J.H. (1990). Agroecology. Intercropping. In Agroecology, C.R. Carrol, J.H. Vandermeer Carroll, and P. Rosset, eds. (New York: McGraw Hill), pp. 481–516.

Vanneste, K., Baele, G., Maere, S., and Van de Peer, Y. (2014). Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. Genome Res. **24**:1334–1347.

Vatanparast, M., Powell, A., Doyle, J.J., and Egan, A.N. (2018). Targeting legume loci: a comparison of three methods for target enrichment bait design in Leguminosae phylogenomics. Appl. Plant Sci. **6**:e1036.

Veizer, J., Godderis, Y., and François, L.M. (2000). Evidence for decoupling of atmospheric CO_2 and global climate during the Phanerozoic eon. Nature **408**:698–701.

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010). KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteomics Bioinformatics **8**:77–80.

Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H., Jin, H., Marler, B., Guo, H., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res **40**, e49–e49.

Wang, J., Sun, P., Li, Y., Liu, Y., Yu, J., Ma, X., Sun, S., Yang, N., Xia, R., Lei, T., et al. (2017). Hierarchically aligning 10 legume genomes establishes a family-level genomics platform. Plant Physiol. **174**:284–300.

Wang, Y.H., Wicke, S., Wang, H., Jin, J.J., Chen, S.Y., Zhang, S.D., Li, D.Z., and Yi, T.S. (2018). Plastid genome evolution in the early-diverging legume subfamily Cercidoideae (Fabaceae). Front. Plant Sci. **9**:138.

Wen, J., Egan, A.N., Dikow, R.B., and Zimmer, E.A. (2015). Utility of transcriptome sequencing for phylogenetic inference and character evolution. In Next-generation Sequencing in Plant Systematics, International Association for Plant Taxonomy, E. Hörandl and M. Appelhans, eds. (Oberreifenberg, Germany: Koeltz Scientic Books), pp. 1–41.

Werner, G.D., Cornwell, W.K., Sprent, J.I., Kattge, J., and Kiers, E.T. (2014). A single evolutionary innovation drives the deep evolution of symbiotic N_2-fixation in angiosperms. Nat. Commun. **5**:4087.

Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M.S., Burleigh, J.G., Gitzendanner, M.A., et al. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc. Natl. Acad. Sci. USA **111**:E4859–E4868.

Wojciechowski, M.F. (2003). Reconstructing the phylogeny of legumes (Leguminosae): an early 21st century perspective. In Higher Level Systematics, Advances in Legume Systematics, B.B. Klitgaard and A. Bruneau, eds. (Kew, UK: Royal Botanic Gardens), pp. 5–35, part 10.

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

Xiang, Y., Huang, C.H., Hu, Y., Wen, J., Li, S., Yi, T., Chen, H., Xiang, J., and Ma, H. (2017). Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. Mol. Biol. Evol. **34**:262–281.

Yang, Y., Moore, M.J., Brockington, S.F., Mikenas, J., Olivieri, J., Walker, J.F., and Smith, S.A. (2018). Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy events in Caryophyllales, including two allopolyploidy events. New Phytol. **217**:855–870.

Yang, Z. (2006). Computational Molecular Evolution (Oxford: Oxford University Press).

Young, N.D., Debellé, F., Oldroyd, G.E.D., Geurts, R., Cannon, S.B., Udvardi, M.K., Benedito, V.A., Mayer, K.F.X., Gouzy, J., Schoof, H., et al. (2011). The *Medicago* genome provides insight into the evolution of rhizobial symbioses. Nature **480**:520–524.

Zachos, J.C., Dickens, G.R., and Zeebe, R.E. (2008). An early Cenozoic perspective on greenhouse warming and carbon-cycle dynamics. Nature **451**:279–283.

Zahran, H.H. (1999). *Rhizobium*-legume symbiosis and nitrogen fixation under severe conditions and in an arid climate. Microbiol. Mol. Biol. Rev. **63**:968–989.

Zeng, L., Zhang, N., Zhang, Q., Endress, P.K., Huang, J., and Ma, H. (2017). Resolution of deep eudicot phylogeny and their temporal diversification using nuclear genes from transcriptomic and genomic datasets. New Phytol. **214**:1338–1354.

Zeng, L., Zhang, Q., Sun, R., Kong, H., Zhang, N., and Ma, H. (2014). Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. Nat. Commun. **5**:4956.

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

Zhang, J., Song, Q., Cregan, P.B., Nelson, R.L., Wang, X., Wu, J., and Jiang, G.L. (2015). Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm. BMC Genomics **16**:217.

Zhang, M., Fritsch, P.W., and Cruz, B.C. (2009). Phylogeny of *Caragana* (Fabaceae) based on DNA sequence data from *rbcL*, *trnS–trnG*, and ITS. Mol. Phylogen. Evol. **50**:547–559.

Zhang, N., Zeng, L., Shan, H., and Ma, H. (2012). Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. New Phytol. **195**:923–937.

Zhang, L., Chen, F., Zhang, X., Li, Z., Zhao, Y., Lohaus, R., Chang, X., Dong, W., Ho, S.Y.W., Liu, X., et al. (2020a). The water lily genome and the early evolution of flowering plants. Nature **577**:79–84.

Zhang, R., Wang, Y.-H., Jin, J.-J., Stull, G.W., Bruneau, A., Cardoso, D., De Queiroz, L.P., Moore, M.J., Zhang, S.-D., Chen, S.-Y., et al. (2020b). Exploration of plastid phylogenomic conflict yields new insights into the deep relationships of Leguminosae. Syst. Biol. **69**:613–622.

Zhang, C., Zhang, T., Luebert, F., Xiang, Y., Huang, C.-H., Hu, Y., Rees, M., Frohlich, M.W., Qi, J., Weigend, M., et al. (2020c). Asterid phylogenomics/phylotranscriptomics uncover morphological evolutionary histories and support phylogenetic placement for numerous whole-genome duplications. Mol. Biol. Evol. **69**. https://doi.org/10.1093/molbev/msaa160.

Zhao, L., Li, X., Zhang, N., Zhang, S.-D., Yi, T.-S., Ma, H., Guo, Z.-H., and Li, D.-Z. (2016). Phylogenomic analyses of large-scale nuclear genes provide new insights into the evolutionary relationships within the rosids. Mol. Phylogen. Evol. **105**:166–176.

Zhuang, W., Chen, H., Yang, M., Wang, J., Pandey, M.K., Zhang, C., Chang, W.-C., Zhang, L., Zhang, X., Tang, R., et al. (2019). The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. Nat. Genet. **51**:865–876.

Zimmerman, E., Herendeen, P.S., Lewis, G.P., and Bruneau, A. (2017). Floral evolution and phylogeny of the Dialioideae, a diverse subfamily of tropical legumes. Am. J. Bot. **104**:1019–1041.