ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202001.00109

Postprint: Phylogenetic Relationships of Angiosperms Based on 5993 Nuclear Genes

Authors: Jin Xin, Cheng' s book, Yang Tuo, Yu Kang, Xiaoxia Duan, Ni Xuemei, Li Shiming, Zhang Gengyun

Date: 2020-01-12T00:00:00+00:00

Abstract

The construction of phylogenetic relationships is crucial for the classification and evolutionary studies of angiosperms. For a long time, studies on angiosperm phylogeny have predominantly utilized plastid genes, mitochondrial genes, or a few conserved single-copy nuclear genes. In this study, we collected nuclear gene sets from 88 angiosperm species (encompassing 58 orders) from annotated genomes or transcriptomes; through orthologous gene clustering and removal of paralogous genes, we obtained 5,993 one-to-one orthologous gene families (i.e., for each gene family, each species had at most one sequence, with a minimum of 50 species included); using DNA or amino acid sequences extracted from various numbers of gene sets, we constructed a total of 20 phylogenetic trees using both concatenation and coalescence methods. Comparing these phylogenetic trees, although most results support the relationships among major angiosperm clades as described in APG IV ((eudicots, monocots), magnoliids), there is a significant difference in the evolutionary relationships among orders within the eudicots compared to APG IV, specifically, this study suggests that Santalales and Caryophyllales are sister groups to the rosids. Based on these phylogenetic trees, we estimated the divergence times of various angiosperm orders, and the results indicate that the origin time of angiosperms was 237.78 million years ago (95% confidence interval: 202.6–278.08 Ma), which is consistent with the mainstream view of 225–240 million years ago. This study provides a feasible strategy for phylogenetic tree construction that allows for the use of a larger number of genes while maintaining faster computational speed.

Full Text

Preamble

DOI: 10.11931/guihaia.gxzw201905048

Title: Reconstruction of Angiosperm Phylogeny Based on 5,993 Nuclear Genes

Authors: JIN Xin[1,2], CHENG Shu[1,2], YANG Tuo[3], YU Kang[1,2], DUAN Xiaoxia[1], , NI Xuemei[1], , LI Shiming[1,2], ZHANG Gengyun[1], *

Affiliations: 1. BGI-Shenzhen, Shenzhen 518083, Guangdong, China 2. BGI Institute of Applied Agriculture, Shenzhen 518120, Guangdong, China 3. China National Gene Bank, Shenzhen 518120, Guangdong, China 4. Key Laboratory of Genomics, Ministry of Agriculture, BGI-Shenzhen, Shenzhen 518083, Guangdong, China

Funding: National Science and Technology Support Program (2015BAD02B01-7); Key Laboratory of Crop Core Resources Development and Application Enterprises of Guangdong (2011A091000047); Science and Technology Program of Shenzhen (JCYJ20150831201123287); Molecular Design and Polymerization Breeding Engineering Laboratory of Shenzhen (Shenfagai[2015]946)

Corresponding Author: ZHANG Gengyun, Ph.D., Researcher. E-mail: zhanggengyun@genomics.cn

---

Abstract

Construction of phylogenetic relationships is crucial for angiosperm classification and evolutionary research. For a long time, angiosperm phylogeny has been analyzed using plastid genes, mitochondrial genes, or a few conserved single-copy nuclear genes. Here, we collected nuclear gene sets from 88 angiosperm species (representing 58 orders) from annotated genomes or transcriptomes. Using a combined homology- and phylogeny-based approach, we obtained a total of 5,993 one-to-one ortholog groups (with at most one sequence per species for each ortholog group), each represented by at least 50 species. We then reconstructed 20 species trees using different combinations of reconstruction methods (concatenation-based and coalescence-based) and sequence types (nucleotide or amino acid) for gene datasets with varying gene occupancy values. Most resulting topologies support the relationships among major angiosperm clades as described in APG IV, but present different deep relationships among major eudicot clades, such as the placement of Santalales and Caryophyllales as sisters to rosids. We estimated the divergence times of major angiosperm clades and concluded that the origin of angiosperms occurred approximately 237.78 million years ago (95% confidence interval: 202.6–278.08 Ma), consistent with the previously accepted timeframe of 225–240 million years ago. This study provides an efficient strategy for building phylogenetic trees using thousands of genes with ultrafast computation.

Keywords: phylogeny, angiosperms, nuclear genes, ortholog coalescence, divergence time inference, concatenation

---

Introduction

Accurate reconstruction of phylogenetic trees is essential for plant classification and evolutionary studies. The accuracy of phylogenetic tree construction is primarily influenced by two factors. First, the type and size of the dataset used: trees built from morphological traits, plastid genes, mitochondrial genes, and nuclear gene sequences differ significantly from one another (Endress & Doyle, 2009; Ruhfel et al., 2014; Soltis et al., 2011; Zeng et al., 2014). Additionally, trees constructed using full-length nucleotide sequences, specific codon positions, or amino acid sequences also show substantial variation (Wickett et al., 2014). Second, the tree-building methods and models employed: the two main approaches are concatenation and coalescence. Concatenation methods treat all genes as a single supermatrix and use software such as RAxML (Stamatakis, 2014) or IQ-TREE (Nguyen et al., 2015) to construct phylogenetic trees. Coalescence methods first build individual gene trees and then use software like ASTRAL (Zhang et al., 2017) to infer a consensus species tree from all gene trees (Wickett et al., 2014). A wide variety of evolutionary models are available, including nucleotide models (GTR, HKY, JC, F81, K2P, K3P, K81uf) and protein models (LG, Poisson, cpREV, mtREV, Dayhoff, mtMAM, JTT, WAG) (Nguyen et al., 2015).

Angiosperms represent the most advanced and diverse group in the plant kingdom, dominating Earth' s vegetation. Currently, 352,000 angiosperm species have been reported (http://www.theplantlist.org/), belonging to 416 families and 64 orders. The evolutionary relationships among these orders have long been a focus of research and debate. Except for the three most basal orders —Amborellales, Nymphaeales, and Austrobaileyales (collectively known as the ANITA grade)—the remaining 99.95% of angiosperms can be divided into five major clades: magnoliids, monocots, eudicots, Chloranthaceae, and Ceratophyllaceae. The phylogenetic topology among these five groups remains controversial. Zeng et al. (2014) summarized five major topologies published to date (Figure 1 [Figure 1: see original paper]: A–E), with Topology A being the most widely accepted and representing the APG IV classification (THE ANGIOSPERM PHYLOGENY GROUP, 2016). Studies using 17 concatenated genes (including plastid, mitochondrial, and nuclear genes) for 640 species (Soltis et al., 2011) and 78 concatenated plastid genes for 360 species (Ruhfel et al., 2014) support Topology A. In contrast, phylogenies based on 674 nuclear genes for 92 species (Wickett et al., 2014) and 59 nuclear genes for 61 species (Zeng et al., 2014) support Topology B. Additionally, Qiu et al. (2010) used four mitochondrial genes for 380 species to support Topology C; Endress & Doyle (2009) used morphological traits to support Topology D; and Zhang et al. (2012) used five nuclear genes for 91 species to support Topology E.

After removing Chloranthaceae and Ceratophyllaceae, three possible relationships exist among magnoliids, monocots, and eudicots: ((eudicots, monocots), magnoliids); ((eudicots, magnoliids), monocots); and ((monocots, magnoliids), eudicots). Lu et al. (2018) analyzed 5,864 Chinese angiosperm species (covering

nearly all Chinese angiosperms) using four plastid genes and one mitochondrial gene, supporting the ((eudicots, monocots), magnoliids) topology. Chen et al. (2019) published the magnoliid *Liriodendron* genome and used 502 nuclear genes with coalescence methods for 18 species, also supporting ((eudicots, monocots), magnoliids). Chaw et al. (2019) published another magnoliid genome (stout camphor tree) and used 211 nuclear genes for 13 species, supporting ((eudicots, magnoliids), monocots). Li et al. (2019) reconstructed a high-resolution angiosperm phylogeny using 80 plastid genes from 2,881 species, supporting ((eudicots, monocots), magnoliids). From these studies, we observe that concatenation methods using nuclear genes generally support ((eudicots, magnoliids), monocots), while coalescence methods using nuclear genes and studies using plastid/mitochondrial genes typically support ((eudicots, monocots), magnoliids).

Phylogenetic relationships within eudicots also remain debated (Figure 1: F–K). Besides the basal orders Ranunculales, Proteales, Trochodendrales, Buxales, and Gunnerales, the remaining eudicots comprise two major clades: rosids and asterids. The phylogenetic relationships among six orders at the base of these clades—Dilleniales, Saxifragales, Vitales, Santalales, Berberidopsidales, and Caryophyllales—are particularly contentious. Zeng et al. (2017) summarized six major topologies (Figure 1: F–K), with Topology K representing the APG IV consensus. Moore et al. (2010) used 83 plastid genes for 86 species to support "Dilleniales as sister to rosids," while Soltis et al. (2011) using 17 concatenated genes and Moore et al. (2011) using plastid IR sequences supported "Dilleniales as sister to asterids." Worberg et al. (2007), Moore et al. (2011), and APG IV supported "Dilleniales as sister to both rosids and asterids."Most studies support "Vitales and Saxifragales as sisters to rosids, and Chilean vine order, Santalales, and Caryophyllales as sisters to asterids" (Moore et al., 2011, 2010; Worberg et al., 2007; Yang et al., 2015). Zeng et al. (2017) used 504 nuclear genes for 100 species to support "Santalales and Berberidopsidales as sisters to rosids."

The origin and evolution of angiosperms have long been hotly debated topics in botany. In paleobotany, the earliest fossil record of angiosperms was long considered to be from the Cretaceous, 125 million years ago, which is also the earliest eudicot fossil record (Herendeen, 1995). Fu et al. (2018) discovered *Nanjinganthus* from Early Jurassic strata (~175 million years ago), which possesses sepals, petals, pistils, a distinct cup-shaped receptacle, epigynous flowers with superior ovaries, and arborescent styles. Its seeds/ovules are completely enclosed, with ovary walls fully isolating seeds from the external environment, meeting all criteria for angiosperm identification. This discovery pushed the earliest angiosperm fossil record back by approximately 50 million years and filled the "Jurassic gap" between the fossil record (125 million years ago) and molecular clock estimates (225–240 million years ago) (Li et al., 2019). Most current molecular dating studies based on phylogenetic trees place angiosperm origins in the Triassic, 225–240 million years ago (Magallon, 2010; Mandel, 2019; Smith et al., 2010; Zeng et al., 2014), consistent with the origin time of pollinating Lepidoptera insects (~230 million years ago) (Li et al., 2019; Zeng et al.,

2014).

This study analyzed the phylogenetic relationships of 88 angiosperm species (representing 87 families and 58 orders) using over 5,000 nuclear genes and both nucleotide and protein sequences with two tree-building methods, and estimated divergence times for all major clades (overall workflow shown in Figure 2 [Figure 2: see original paper]). To obtain accurate and reliable angiosperm phylogenies, we partitioned the >5,000 nuclear genes into multiple datasets containing different numbers of genes, constructed phylogenetic trees from each dataset, and compared the consistency among the resulting 20 phylogenetic trees.

---

1. Materials and Methods

1.1 Materials

We collected one gymnosperm genome (*Ginkgo biloba* as outgroup), 43 angiosperm genomes (primarily from NCBI and Phytozome databases), 43 assembled angiosperm transcriptomes (http://www.onekp.com/public_data.html), and two angiosperm RNA-seq datasets (including *Petrosavia sakurai*, sequenced in this study). The angiosperm samples comprised 87 families across 58 orders (Table 1).

1.2 Ortholog Identification from Genomic Sequences

We performed homologous gene clustering on gene sets from 43 plant genomes using the method reported by Yang & Smith (2014). First, we conducted all-by-all BLASTN v2.6.0+ comparisons on CDS sequences from the 43 gene sets, retaining the top 1,000 hits per query. We removed alignments shorter than one-third of the total sequence length and trimmed unaligned terminal regions. We then used MCL software (Van, 2000) for homologous gene clustering (inflation value = 1.4), removed gene families with fewer than 20 species, and performed multiple sequence alignment using MAFFT v7.310 (Katoh & Standley, 2013) with maximum iterative refinement cycles set to 1,000. We trimmed sites with >90% missing data using PHYLIP v2.2.6 (Smith & Dunn, 2008) and estimated phylogenetic trees using RAxML v8.2.11 (Stamatakis, 2014) with the GTRCAT model. Finally, we removed all paralogous branches from the resulting trees, including branches >0.6 in length, terminal branches ten times longer than their sister branches, and all but one branch from monophyletic groups composed entirely of the same sample. We also trimmed internal branches with lengths exceeding the expected substitution rate by 0.3, then applied the MO method (Yang & Smith, 2014) to remove any remaining paralogous branches, yielding one-to-one ortholog families (at most one sequence per sample). Only gene families with >20 samples were retained.

1.3 Transcriptome and Outgroup Data Processing

We de novo assembled RNA-seq data from two families (*Petrosavia sakurai* from Petrosaviaceae and *Cyanotis arachnoidea* from Commelinaceae). First, we filtered raw reads using Trimmomatic v0.38 (Bolger et al., 2014) with parameters: HEADCROP:15 LEADING:20 TRAILING:20 SLIDINGWINDOW:5:20 MINLEN:50 AVGQUAL:20. We then assembled transcripts using Trinity v2.6.6 (Grabherr et al., 2011) (minimum contig length = 150 bp) and predicted CDS and protein sequences using TransDecoder v5.5.0 (https://github.com/TransDecoder/TransDecoder/releases/tag/TransDecoder-v5.5.0) with Swissprot and Pfam-A as reference databases. We merged gene sets from these two species, 43 angiosperm transcriptomes from the 1KP database, and one gymnosperm (*Ginkgo biloba*) into the ortholog families obtained from genomic data using HaMStR v13.2.6 (Ebersberger et al., 2009), retaining only gene families with >50 samples.

1.4 Phylogenetic Tree Construction

We employed two methods—concatenation and coalescence—using both CDS and amino acid sequences. Both sequence types were aligned using PRANK v.170427 (http://wasabiapp.org/software/prank/), trimmed for sites with >70% missing data using PHYLIP v2.2.6 (Smith & Dunn, 2008), and filtered for length (CDS sequences <300 bp and protein sequences <100 amino acids were removed).

Coalescence method: We first constructed individual gene trees using RAxML v8.2.11 (default parameters) (Stamatakis, 2014), then processed all gene trees using ASTRAL v5.5.9 (Zhang et al., 2017) to obtain a consensus tree. Parameters were set to "-t 1 -gene-only" to obtain bootstrap values and gene support scores. Branch lengths were obtained using IQ-TREE v1.5.5 (Nguyen et al., 2015).

Concatenation method: We used PartitionFinder v2.1.1 (Lanfear et al., 2009) to detect optimal partitioning schemes and evolutionary models for the concatenated sequences. For CDS sequences, we tested four partitioning strategies (Table 2): no partitioning, partitioning by codon position (three partitions), partitioning by gene, and partitioning by codon position within each gene. For protein sequences, we tested two strategies: no partitioning and partitioning by gene. Parameters were set as: branch lengths = linked; model_selection = aicc; search = user; models = GTR, GTR+G, GTR+I+G (for CDS) or models = LG+G, LG+I+G, WAG+G, WAG+I+G (for proteins). We then constructed trees using IQ-TREE v1.5.5 (1,000 ultrafast bootstrap replicates (Von Haeseler et al., 2013), with "-spp" specifying the optimal partitioning scheme). Gene support values were obtained using ASTRAL v5.5.9 (-t 1). All trees were visualized using Evolview v2 (He et al., 2016).

1.5 Divergence Time Estimation

We estimated divergence times using the MCMCTREE program in PAML v4.9 (Yang, 2007). The input topology was the best topology from our 20 trees (concatenation-based tree using CDS sequences from 742 genes), and input sequences were CDS sequences from the 742 genes. We first estimated divergence times for each gene separately, then integrated results across all 742 genes (taking the mean value for each node) to obtain the final chronogram.

Branch lengths were obtained using the JONES+gamma substitution model; rgene gamma was set to G(1, 4.5); sigma2 gamma was set to G(1, 4.5); clock was set to 3; MCMC parameters were burnin = 50,000, sampfreq = 100, nsample = 10,000. For each gene, we ran two independent MCMC chains (with different random seeds) and assessed convergence using Tracer v1.7 (https://github.com/beast-dev/tracer/releases/tag/v1.7.1), ensuring all nodes and parameters had effective sample sizes >200. Nine fossil calibration points were used: *Ginkgo biloba* divergence at 290–310 Ma (Gao et al., 1989); monocot-eudicot divergence at 130–200 Ma (Kumar et al., 2017); eudicot crown age (earliest eudicot fossil record) at 125 Ma (Herendeen, 1995; Zeng et al., 2014); Proteales crown at 108.8 Ma (Crane et al., 1996); Vitales-rosids divergence at 105–115 Ma (Fawcett et al., 2009; Kumar et al., 2017); *A. thaliana*–*P. trichocarpa* divergence at 97–109 Ma (Kumar et al., 2017); Fabales-Fagales divergence at 93.5 Ma (Friis et al., 1996); Cornales crown at 85.8 Ma (Takahashi et al., 2002); and Lamiales crown at 44.3 Ma (Call et al., 1992).

---

2. Results

2.1 Ortholog Identification

We performed homologous gene clustering on CDS sequences from 44 plant genomes and 45 assembled transcriptomes, applying the paralog removal method of Yang & Smith (2014) to obtain 5,993 one-to-one ortholog families with >50 samples (Figure 3 [Figure 3: see original paper]:A). Gene coverage across species ranged from 33.57% to 97.85%, with an average of 80.40% (Figure 3:B).

2.2 Phylogenetic Tree Construction

We constructed 20 phylogenetic trees using concatenation and coalescence methods to assess topological stability (Figure 4 [Figure 4: see original paper]). Both CDS and protein sequences were analyzed using five datasets each, yielding 20 total trees (5 CDS concatenation, 5 CDS coalescence, 5 protein concatenation, and 5 protein coalescence). The five datasets contained 5,928 orthologs (50 samples), 3,384 orthologs (70 samples), 1,791 orthologs (80 samples), 742 orthologs (85 samples), and 42 orthologs (89 samples).

These 20 trees primarily aimed to resolve relationships among the five major

angiosperm clades and relationships among orders within eudicots. Most trees were highly consistent with the concatenation-based tree constructed using CDS sequences from 742 genes (4,069,848 sites) (Figure 5 [Figure 5: see original paper]) (the concatenation-based trees using 3,384 and 1,791 protein genes also yielded this optimal topology).

2.2.1 Relationships Among Magnoliids, Monocots, and Eudicots Regardless of sequence type (nucleotide or protein) or method (concatenation or coalescence), most trees supported the topology ((eudicots, monocots), magnoliids) (Figure 4).

2.2.2 Chloranthaceae and Ceratophyllaceae Our study identified Ceratophyllaceae as sister to eudicots, consistent with previous research (Figure 4). However, Chloranthaceae was resolved as the basal sister to all angiosperms except the ANITA grade, differing from APG IV' s placement of Chloranthaceae as sister to magnoliids.

2.2.3 Relationships Among Orders Within Eudicots Our results support Dilleniaceae as sister to both rosids and asterids, and Saxifragales as sister to rosids, both consistent with APG IV (Figure 4).

APG IV proposes that "Santalales and Caryophyllales are sisters to asterids," but our study rejects this conclusion. All 20 trees support "Caryophyllales as sister to rosids." Most trees support "Santalales as sister to rosids," consistent with Zeng et al. (2017), while a minority support "Santalales as sister to both rosids and asterids" (Figure 4).

APG IV considers "Berberidopsidales as sister to asterids,"but only a minority of our trees support this. Protein-based trees (both concatenation and coalescence) support "Berberidopsidales as sister to both rosids and asterids,"while nucleotide-based trees increasingly support "Berberidopsidales as sister to asterids" as gene number increases, aligning with APG IV (Figure 4).

2.3 Divergence Time Estimation

Based on the concatenation tree from 742-gene CDS sequences, we estimated angiosperm divergence times (Figure 6 [Figure 6: see original paper]). We estimate the origin of angiosperms at 237.78 million years ago (95% CI: 202.6–278.08 Ma), consistent with the mainstream view of 225–240 million years ago (Magallon, 2010; Smith et al., 2010; Zeng et al., 2014). The divergence between magnoliids and the monocot-eudicot lineage occurred approximately 166.11 Ma; Dilleniaceae diverged from rosids and asterids at ~124.23 Ma; rosids and asterids diverged at ~116.98 Ma; and lamiids and campanulids diverged at ~102.37 Ma.

3. Discussion and Conclusion

Traditionally, angiosperm phylogenetic reconstruction has relied on plastid genes, mitochondrial genes, or a few conserved single-copy nuclear genes. Yang & Smith (2014) reported a phylogeny-based method for homologous gene clustering and paralog removal. We applied this approach to nuclear gene sets from 88 plant species, obtaining 5,993 one-to-one ortholog families. We then subsampled this dataset to create various-sized subsets for tree reconstruction to assess stability.

After obtaining a large nuclear gene dataset, computational resources and time become limiting factors. Phylogenetic construction typically requires bootstrap resampling (100–1,000 replicates), which is computationally intensive. Nguyen et al. (2015) introduced IQ-TREE with ultrafast bootstrap approximation (UF-Boot), which is 10–40 times faster than traditional RAxML methods while providing more accurate bootstrap values (Von Haeseler et al., 2013).

Our phylogenies based on 5,993 one-to-one ortholog families differ from APG IV primarily in the placement of Santalales and Caryophyllales. Our study places both orders as sisters to rosids, whereas APG IV positions them as sisters to asterids. Two potential reasons may explain this discrepancy: first, the substantial increase in gene number; second, our dataset comprises approximately equal numbers of genome and transcriptome sequences, with transcriptomes typically exhibiting substantial gene missingness (i.e., many unexpressed genes).

Overall, this study not only clarifies phylogenetic relationships among angiosperm orders but also explores a feasible strategy for building trees with more genes and faster computation: using the homolog clustering and paralog removal method of Yang & Smith (2014) to obtain numerous one-to-one ortholog families, followed by rapid and accurate tree construction with IQ-TREE (concatenation) and ASTRAL (coalescence). As more plant genomes are sequenced and gene clustering and phylogenetic methods continue to improve, angiosperm phylogenetic relationships will become increasingly resolved, enabling more precise determination of the relationships of Santalales and Caryophyllales with other angiosperm lineages.

References

Bolger AM, Lohse M, Usadel B, 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15): 2114-2120.

Call VB, Dilcher DL, 1992. Investigations of angiosperms from the Eocene of southeastern North America: Samaras of *Fraxinus wilcoxiana* Berry. *Rev Palaeobot Palynol*, 74: 249-266.

Chaw SM, Liu YC, Wu YW, et al., 2019. Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution. *Nat Plants*, 5(1): 63-73.

Chen JH, Hao ZD, Guang XM, et al., 2019. *Liriodendron* genome sheds light on angiosperm phylogeny and species-pair differentiation. *Nat Plants*, 5(1): 18-25.

Crane PR, Herendeen PS, 1996. Cretaceous floras containing angiosperm flowers and fruits from eastern North America. *Rev Palaeobot Palynol*, 90: 319-337.

Ebersberger I, Strauss S, Von Haeseler A, 2009. HaMStR: Profile hidden markov model based search for orthologs in ESTs. *Bmc Evol Biol*, 9(1): 157-157.

Endress PK, Doyle JA, 2009. Reconstructing the ancestral angiosperm flower and its initial specializations. *Amer J Bot*, 96(1): 22-66.

Fawcett JA, Maere S, Van De Peer Y, 2009. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc Nat Acad Sci USA*, 106(14): 5737-5742.

Friis EM, Pedersen KR, Schönenberger J, 2006. Normapolles plants: A prominent component of the Cretaceous rosid diversification. *Plant Syst Evol*, 260: 107-140.

Fu Q, Diez JB, Pole M, et al., 2018. An unexpected noncarpellate epigynous flower from the Jurassic of China. *Elife*, 7: e38827.

Gao Z, Barry AT, 1989. A review of fossil cycad megasporophylls, with new evidence of *Crossozamia pomel* and its associated leaves from the lower permian of Taiyuan, China. *REV Palaeobot Palynol*, 60(3-4): 205-223.

Grabherr MG, Haas BJ, Yassour M, et al., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*, 29(7): 644-652.

He ZL, Zhang HK, Gao SH, et al., 2016. Evolview v2: An online visualization and management tool for customized and annotated phylogenetic trees. *Nucl Acid Res*, 44(W1): W236-W241.

Herendeen PS, 1995. The enigma of angiosperm origins. *Earth-Sci Rev*, 39(1): 253-254.

Katoh K, Standley DM, 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol*, 30(4): 772-780.

Kumar S, Stecher G, Suleski M, et al., 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol*, 34: 1812-1819.

Lanfear R, Frandsen PB, Wright AM, et al., 2016. PartitionFinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol Biol Evol*, 34(3): 772-773.

Li HT, Yi TS, Gao LM, et al., 2019. Origin of angiosperms and the puzzle of the Jurassic gap. *Nat Plants*, 5(1): 461-470.

Lu LM, Mao LF, Yang T, et al., 2018. Evolutionary history of the angiosperm flora of China. *Nature*, 554(1): 234-238.

Magallon S, 2010. Using fossils to break long branches in molecular dating: A comparison of relaxed clocks applied to the origin of angiosperms. *Syst Biol*, 59(4): 384-399.

Moore MJ, Hassan N, Gitzendanner MA, et al., 2011. Phylogenetic Analysis of the Plastid Inverted Repeat for 244 Species: Insights into Deeper-Level Angiosperm Relationships from a Long, Slowly Evolving Sequence Region. *Int J Plant Sci*, 172(4): 541-558.

Moore MJ, Soltis PS, Bell CD, et al., 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proc Nat Acad Sci USA*, 107(10): 4623-4628.

Nguyen LT, Schmidt HA, Von Haeseler A, et al., 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*, 32(1): 268-274.

Qiu YL, Li LB, Wang B, et al., 2010. Angiosperm phylogeny inferred from sequences of four mitochondrial genes. *JSE*, 48(6): 391-425.

Ruhfel BR, Gitzendanner MA, Soltis PS, et al., 2014. From algae to angiosperms–inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *Bmc Evol Biol*, 14(1): 23.

Smith SA, Beaulieu JM, Donoghue MJ, 2010. An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. *Proc Nat Acad Sci USA*, 107(13): 5897-5902.

Smith SA, Dunn CW, 2008. Phyutility: A phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics*, 24(5): 715-716.

Soltis DE, Smith SA, Cellinese N, et al., 2011. Angiosperm phylogeny: 17 genes, 640 taxa. *Amer J Bot*, 98(4): 704-730.

Stamatakis A, 2014. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics*, 30(9): 1312-1313.

Takahashi M, Crane PR, Manchester SR, 2002. *Hironoia fusiformis* gen. et sp. nov., A cornalean fruit from the Kamikitaba locality (Upper Cretaceous, Lower Coniacian) in northeastern Japan. *J Plant Res*, 115: 463-473.

THE ANGIOSPERM PHYLOGENY GROUP, 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot J Linn Soc*, 181(1): 1-20.

Van DS, 2000. *Graph Clustering by Flow Simulation*. University of Utrecht.

Von Haeseler A, Minh BQ, Nguyen MAT, 2013. Ultrafast Approximation for Phylogenetic Bootstrap. *Mol Biol Evol*, 30(5): 1188-1195.

Wickett NJ, Mirarab S, Nguyen N, et al., 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc Nat Acad Sci USA*, 111(45): 4859-4868.

Worberg A, Quandt D, Barniske AM, et al., 2007. Phylogeny of basal eudicots: Insights from non-coding and rapidly evolving DNA. *ORG DIVERS EVOL*, 7(1): 55-77.

Yang Z, 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, 24: 1586–1591.

Yang Y, Moore MJ, Brockington SF, et al., 2015. Dissecting Molecular Evolution in the Highly Diverse Plant Clade Caryophyllales Using Transcriptome Sequencing. *Mol Biol Evol*, 32(8): 2001-2014.

Yang Y, Smith SA, 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: Improving accuracy and matrix occupancy for phylogenomics. *Mol Biol Evol*, 31(11): 3081-3092.

Zeng LP, Zhang N, Zhang Q, et al., 2017. Resolution of deep eudicot phylogeny and their temporal diversification using nuclear genes from transcriptomic and genomic datasets. *New Phytol*, 214(3): 1338-1354.

Zeng LP, Zhang Q, Sun RR, et al., 2014. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nat Comm*, 5(1): 4956.

Zhang C, Sayyari E, Mirarab S, 2017. ASTRAL-III: Increased Scalability and Impacts of Contracting Low Support Branches. *RECOMB-CG*, Springer, Cham: 53-75.

Zhang N, Zeng LP, Shan HY, et al., 2012. Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. *New Phytol*, 195(4): 923-937.

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv —Machine translation. Verify with original.