


Evolution of the lignin biosynthesis pathway underlies life form strategy diversification in Rosaceae

Li-Yao Su, Zheng-Tai Liu, Pei-Yan Chen, Xi-Liang Wang, Jin-Song Xiong and Ai-Sheng Xiong* 

State Key Laboratory of Crop Genetics & Germplasm Enhancement and Utilization, Ministry of Agriculture and Rural Affairs Key Laboratory of Biology and Germplasm Enhancement of Horticultural Crops in East China, College of Horticulture, Nanjing Agricultural University, Nanjing 210095, China

* Corresponding author, E-mail: xiongaisheng@njau.edu.cn

Abstract

As a key component of the plant secondary cell wall, lignin is crucial for mechanical support and water transport. The Rosaceae family, with its diverse life forms, is an ideal group for studying the relationship between lignin biosynthesis and life form evolution. Based on genomic data from 37 Rosaceae species, this study reconstructed the phylogenetic history and analyzed gene family expansion and contraction. Results show that woody lineages experienced significant expansion of gene families related to lignin biosynthesis at key evolutionary nodes, while herbaceous plants showed a trend of contraction. Specifically, the *CAD*, *COMT*, and *HCT* gene families notably increased in copy number in woody plants, primarily through whole-genome duplication and tandem duplication events. Phylogenetic analysis revealed a woody plant-specific clade in the *HCT* family, suggesting functional specialization after duplication. Some gene clusters exhibited higher Ka/Ks ratios, indicating potential positive selection. This study demonstrates that the expansion and functional differentiation of lignin biosynthesis genes are key molecular mechanisms driving life form diversification in Rosaceae, providing new insights into the evolution of plant mechanical traits.

Citation: Su LY, Liu ZT, Chen PY, Wang XL, Xiong JS, et al. 2025. Evolution of the lignin biosynthesis pathway underlies life form strategy diversification in Rosaceae. *Genomics Communications* 2: e025 <https://doi.org/10.48130/gcomm-0025-0025>

Introduction

Based on the degree of lignification, plant life forms are classified into herbs, shrubs, and trees. Plants of different life forms exhibit distinct responses to external environmental changes. Due to significant ecological differences between herbaceous and woody plants in terms of adaptation to abiotic environments, functional traits, and dispersal capacity, their spatial patterns of species diversity also vary accordingly^[1–3]. These differences highlight that the community assembly mechanisms may vary across plant life forms. For instance, in tropical forests, herbaceous plant communities exhibit a stronger environmental filtering effect compared to woody plant communities^[1,4]. Rosaceae is a highly distinctive and significant plant family, encompassing a vast number of edible fruit trees, ornamental flowers, and medicinal plants. The remarkable diversification within Rosaceae makes it an excellent group for studying plant evolution. The diversity of its fruit types provides exceptional research materials for investigating the formation and development of plant fruits^[5]. Similarly, the life forms of Rosaceae plants are highly diverse. The genus *Dryas* is currently recognized as the basal lineage within the family^[5]. The common ancestor of Rosaceae subsequently diversified into herbs (e.g., strawberry, cinquefoil, *Gillenia trifoliata*), shrubs (e.g., rose, rugosa rose, raspberry), and trees (e.g., apple, pear, peach) during its evolutionary history. This remarkable diversity is the result of long-term evolution, yet the underlying genetic mechanisms remain unclear.

Lignin, the second most abundant polymer on earth, is a product of the phenylpropanoid biosynthesis pathway and serves as one of the most critical metabolites for plant mechanical properties^[6]. It plays a fundamental role in plant growth and development. Beyond development, lignin also fulfills important functions in plant responses to both biotic and abiotic stresses^[7–9]. Naturally occurring lignin polymers are composed of three basic monomeric units: p-hydroxyphenyl (H), guaiacyl (G), and syringyl (S) (Lignification:

Flexibility, Biosynthesis and Regulation)^[10]. The biosynthesis of lignin primarily involves the following steps: Plants utilize phenylalanine as a substrate, which enters the phenylpropanoid pathway catalyzed by enzymes encoded by genes such as *PAL* (phenylalanine ammonia lyase), *C4H* (cinnamate-4-hydroxylase), and *4CL* (4-coumarate: CoA ligase), leading to the generation of universal precursor compounds. Subsequently, through the regulation of genes including *HCT* (hydroxycinnamoyl-CoA:shikimate/quinic acid hydroxycinnamoyltransferase), *COMT* (caffeic acid 3-O-methyltransferase), and *CCoAOMT* (trans-caffeoyl-CoA 3-O-methyltransferase), the three lignin monomers (G-type, S-type, and H-type) are synthesized. Finally, these monomers are oxidatively polymerized by enzymes such as peroxidases (*PRX*) and laccases (*LAC*), forming the complex lignin polymer. Therefore, evolutionary changes in plant life forms are closely linked to the biosynthesis of lignin^[11].

Lignin is a key component of plant cell walls, with its primary functions including enhancing mechanical strength, providing protection against pathogens, and regulating water transport pathways. Significant variations in lignin content and composition exist among plants of different life forms. For instance, woody plants generally exhibit higher lignin content, which contributes to the formation of robust lignified cell walls that offer structural support, enabling adaptation to long-term survival in complex environments such as those involving wind and rain. In contrast, herbaceous plants typically possess lower lignin content, favoring flexibility to accommodate rapid growth and development, as well as dynamic environmental changes^[12,13]. Studies have also shown that aquatic plants often display further reduced lignin content, thereby conserving energy and resources to suit environments with low mechanical strength demands and submerged growth conditions^[14]. The relationship between lignin content and plant life form is particularly evident in adaptive evolution. Short-lived plants, regardless of their habitat, tend to have lower lignin content to support the

completion of a rapid life cycle, whereas long-lived perennials or woody plants generally exhibit higher lignin content to enhance mechanical support and durability^[14,15].

In previous research, studies on lignin biosynthesis have primarily focused on single species or individual gene families. Investigations into gene families related to lignin biosynthesis have been conducted in crops such as rice^[16], wheat^[17], tea^[18], citrus^[19], and walnut^[20]. As one of the most critical factors in plant terrestrialization, the origin and evolution of lignin biosynthesis during this process have been examined, providing key evidence for understanding the mechanisms underlying the formation of plant mechanical traits^[21]. However, these previous studies have generally been conducted from a broad perspective across the plant kingdom, lacking in-depth exploration within specific plant lineages. To address this knowledge gap and directly investigate the role of lignin biosynthesis in the diversification of plant life forms, this study conducted comparative genomic analyses using the Rosaceae family, which encompasses herbaceous plants, shrubs, and trees, as an ideal model. It was systematically revealed, for the first time, that key gene families such as *CAD*, *COMT*, and *HCT* have contributed to the evolution of lignification traits in Rosaceae through lineage-specific expansions and functional divergence. These findings provide novel and critical molecular evidence for understanding the genetic basis underlying the diversification of plant life forms.

Materials and methods

Construction of a phylogenetic tree and identification of key evolutionary nodes for 37 Rosaceae species

The genomic data for the 37 Rosaceae species used in this study were obtained from the Genome Database for Rosaceae (GDR, www.rosaceae.org), and other publicly available genomic databases (Supplementary Table S1). Protein sequences of the 37 species were clustered using OrthoFinder v3.1^[22]. Based on the OrthoFinder results, only nine single-copy gene families were identified among the 37 species. Therefore, 274 low-copy gene families with a copy number of less than three were selected, and one member from each family was randomly chosen to infer phylogenetic trees using both concatenation and coalescent-based methods. To estimate the divergence times across the Rosaceae family, fossil calibrations from the Rosoideae (33.2 MYA) and Amygdaloideae (43.2 MYA) subfamilies were used. Subsequently, based on the evolutionary relationships in the phylogenetic tree, expanded and contracted gene families for each species and key evolutionary nodes were identified and analyzed using CAFÉ5^[23]. Finally, members of the expanded and contracted gene families at key nodes were extracted and subjected to Gene Ontology (GO) enrichment analysis using the clusterProfiler software^[24].

Identification of genes in the lignin biosynthesis pathway

Using the known lignin biosynthesis gene members from *Arabidopsis thaliana* as query sequences (Supplementary Table S2), a BLASTP^[25] analysis was performed against the protein sequences of the 37 Rosaceae species with stringent parameters, setting an e-value cutoff of 1e-20. The resulting hits were subsequently annotated for conserved domains using the InterProScan software^[26]. The databases employed in this study were Pfam and PANTHER. Based on the identification of these putative lignin biosynthesis gene members, the complete set of gene sequences involved in lignin biosynthesis was obtained. A preliminary phylogenetic tree was constructed using FastTree v 2.1.11^[27], and members with anomalous branch lengths were subsequently removed.

Construction of syntenic networks and Ka/Ks analysis

To elucidate the genomic evolutionary history (e.g., duplication events and ancestral relationships) of the *CAD*, *COMT*, and *HCT* families, intra- and inter-species synteny analyses were performed for all identified gene members across the 37 Rosaceae genomes using JCVI (Python version of MCSan)^[28]. Two genes were considered syntenic if they were located within conserved genomic blocks. A syntenic network was then constructed for each gene family, where nodes represent individual genes and connecting edges represent detected syntenic relationships. Subsequently, tandemly and proximally duplicated gene pairs were identified in the 37 plant genomes using DupGen_finder^[29]. The identified gene pairs were then aligned, and their sequence alignment formats were converted using ParaAT v2.0^[30]. Finally, the Ka/Ks values for all gene pairs were calculated using KaKs_Calculator 3.0^[31].

Phylogenetic analysis of CAD, COMT, and HCT proteins

The protein sequences (*A. thaliana* and 37 Rosaceae species) of the identified CAD (Cinnamyl alcohol dehydrogenase), COMT (Caffeic acid 3-O-methyltransferase), and HCT (Hydroxycinnamoyl-CoA:shikimate/ quinate hydroxycinnamoyltransferase) gene family members were aligned using MAFFT v7.505^[32]. The resulting multiple sequence alignments were then processed with Phyutility software using the -clean 0.5 parameter to trim gap-rich regions^[33]. Subsequently, the best-fit model for phylogenetic tree construction was selected using ProtTest v3.4.2 with the parameters -all-distributions -F -AIC -BIC -tc 0.5^[34]. Finally, phylogenetic trees for the CAD, COMT, and HCT protein members were constructed separately using IQ-TREE v2.0 with the JTT + G + F, JTT + I + G + F, and JTT + G + F models, respectively, with 1,000 bootstrap replicates^[35]. Conserved motifs within the CAD, COMT, and HCT members were identified using MEME v5.5.8 with the following parameters: maximum number of motifs = 15, minimum width = 6, maximum width = 50^[36]. Finally, the frequency of each motif in different branches of the gene family was calculated by dividing the occurrence count of each motif by the total number of genes in the corresponding branch.

Results

Phylogenetic framework and lineage-specific gene family dynamics in Rosaceae

The Rosaceae family encompasses a wide range of herbs, shrubs, and trees. To investigate the evolutionary mechanisms behind this diversification, this study collected published genomic data from 37 Rosaceae species, most of which are at the chromosome level. The selected species include *Dryas drummondii* from the basal subfamily Dryadoideae, 13 species from Rosoideae (including common species such as strawberry, rose, and *Rosa rugosa*), and 23 species from Amygdaloideae (including common species such as apple, pear, and peach). A phylogenetic tree was constructed based on 274 low-copy gene families from these 37 species, and gene family expansion and contraction were calculated for each species (Fig. 1). The results indicate that Rosoideae and Amygdaloideae diverged approximately 83.82 million years ago. During subsequent evolution, *Rubus* and *Rosa* successively diverged from *Fragaria*, while *Prunus*, *Malus*, and *Pyrus* also diverged within Amygdaloideae (Supplementary Fig. S1). Therefore, we hypothesize that the extensive gene gains and losses during the diversification of Rosaceae may have contributed to the differentiation into herbs, shrubs, and trees. Accordingly, 11 key evolutionary nodes were annotated (A–K) on the species tree of the 37 plants to identify changes in gene families at these points.

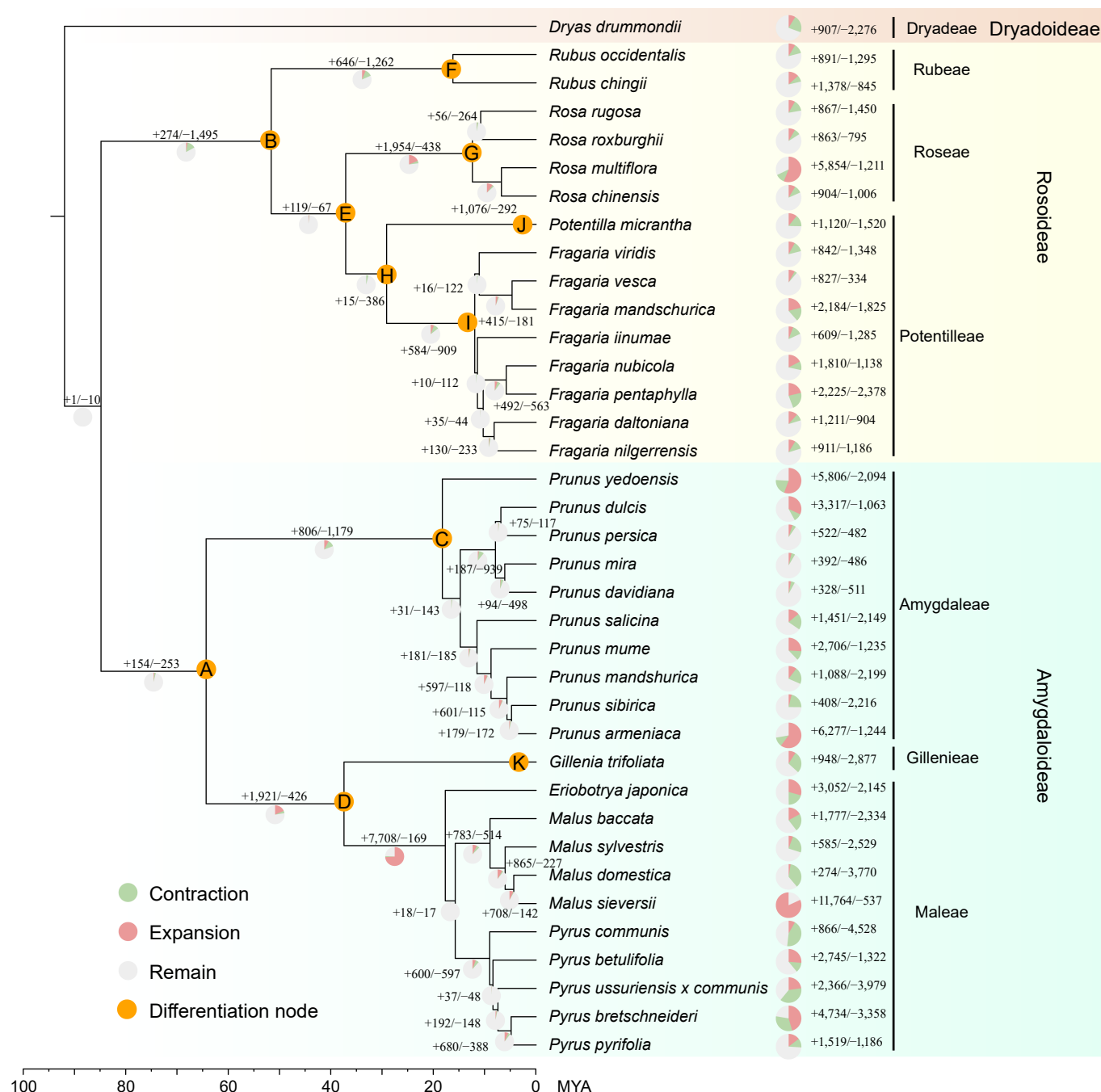


Fig. 1 Phylogeny and gene family expansion/contraction analysis in 37 Rosaceae species.

Furthermore, expanded and contracted gene families at nodes A–K were extracted for Gene Ontology (GO) enrichment analysis. The results revealed that a substantial number of genes involved in pathways such as cellulose synthesis, lignin biosynthesis, and cell wall biosynthesis were present at these eleven key phylogenetic nodes (Fig. 2). Within the highly lignified Amygdaloideae, four key nodes were marked: A, C, D, and K. The analysis showed that nodes A and D both exhibited significant expansion of gene families associated with cellulose synthesis, lignin biosynthesis, and cell wall biosynthesis pathways. In contrast, nodes C and K showed contraction of gene families in these same pathways (Fig. 2). In Rosoideae, seven nodes were annotated: B, E, F, G, H, I, and J. Among these, only nodes E, F, and G demonstrated expansion of gene families related to cellulose synthesis, lignin biosynthesis, and cell wall

biosynthesis pathways. The remaining nodes exhibited contraction of gene families in these biosynthetic pathways (Fig. 2). In summary, the findings of this study indicate that herbaceous plants lost the most gene families associated with cellulose synthesis, lignin biosynthesis, and cell wall biosynthesis during evolution, followed by shrubs, with woody plants showing the fewest losses.

Lignin biosynthesis gene copy number variation correlates with life form diversification

As the primary pathway for synthesizing lignin monomers in plants, understanding the lignin biosynthesis pathway in Rosaceae species can provide valuable insights into the diversification within this family. This study first identified 10 genes involved in the lignin biosynthesis pathway and illustrated the schematic diagram of lignin biosynthesis in plants (Fig. 3a). Using the basal Rosaceae species *Dryas drummondii*

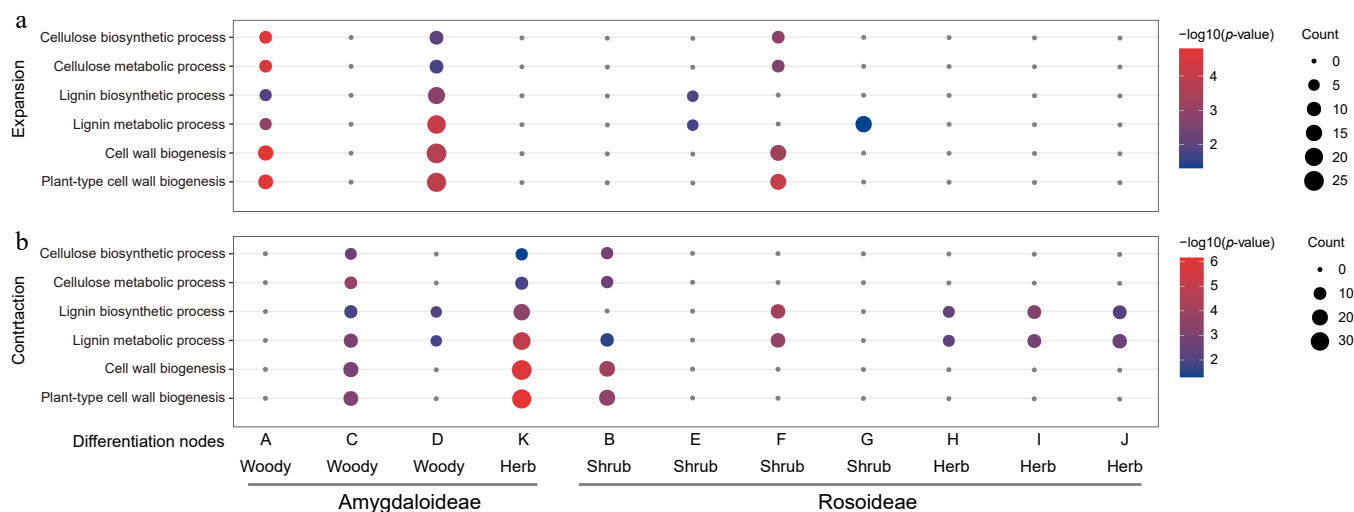


Fig. 2 GO enrichment analysis of (a) expanded, and (b) contracted gene families at 11 key evolutionary nodes on the Rosaceae phylogenetic tree.

(herbaceous) as a reference, a significant increase in the copy numbers of *CAD* (Cinnamyl alcohol dehydrogenase), *COMT* (Caffeic acid 3-O-methyltransferase), and *HCT* (Hydroxycinnamoyl-CoA:shikimate/quinic acid hydroxycinnamoyltransferase) gene family members was observed in the *Rubus* and *Rosa* plants within the Rosoideae clade. In contrast, the number of gene family members in *Fragaria* showed no significant change compared to *Dryas drummondii* (Fig. 3b). Within the Amygdaloideae clade, the number of gene families associated with lignin biosynthesis generally increased, with nearly all gene families exhibiting a significant expansion in copy number in *Malus* and *Pyrus* (Fig. 3b). However, an exception was noted in *Gillenia trifoliata*, a unique species within Amygdaloideae. Despite sharing a common ancestor and key expansion nodes with *Malus* and *Pyrus*, the number of lignin biosynthesis-related gene families in *Gillenia trifoliata* remained largely similar to that in *Dryas drummondii* (Fig. 3b). In summary, compared to herbaceous plants, both shrubs and trees exhibited an overall increase in the absolute copy number of lignin-related gene families.

Further investigation into the gene duplication events responsible for these additional gene copies revealed distinct patterns among herbaceous, shrub, and woody plants. The *CAD*, *COMT*, and *HCT* gene families exhibited the highest numbers of whole-genome duplications (WGD), tandem duplications (TD), and proximal duplications (PD), respectively, indicating that these three gene families may play critical roles in the development of woody traits in Rosaceae. Additionally, the Ks values of WGD-derived gene pairs in different Rosaceae life forms were distributed within the range of 2–4. While herbaceous and shrub plants showed distributions primarily within this range, their numbers were relatively low. In contrast, woody plants such as apple and pear possessed a substantial number of WGD-derived gene pairs with Ks values distributed between 0–1, which is associated with polyploidization events (Fig. 3c). Most TD- and PD-derived gene pairs had Ks values distributed between 0–1, with herbaceous plants exhibiting the fewest duplication events (Fig. 3d, e). The degree of lignification in plants increased with the rising number of gene copies, suggesting that the origin of woody traits in Rosaceae is linked to genome duplication and additional gene duplication events.

The *CAD* gene family expanded through duplication and underwent positive selection in woody lineages

Based on the aforementioned research, a total of 790 *CAD* gene family members were identified from the 37 plant species. A phylogenetic

analysis was subsequently conducted by integrating these 790 members with nine *CAD* members from *A. thaliana*. The results demonstrated that the *CAD* members in Rosaceae can be divided into six distinct clades. These include clades shared with *Arabidopsis*, namely CAD1, CAD2/3/9, CAD4/5, CAD6/7/8, and two clades that contained only Rosaceae sequences (hereafter referred to as Rosaceae-lineage-specific clades) (Fig. 4a). To investigate functional divergence among different clades, the frequency of 15 conserved motifs was analyzed. The results indicated that, with the exception of motif 15, the occurrence frequencies of the other motifs were largely consistent across clades. The significantly lower frequency of motif 15 in the CAD1 and CAD4/5 clades suggesting potential structural and functional divergence among these clades (Fig. 4b). Furthermore, a syntenic network of the *CAD* gene family in Rosaceae was constructed. The network revealed that the *CAD* genes can be grouped into four major gene clusters (Fig. 4c), indicating that the Rosaceae *CAD* family may have originated from four ancestral genes. Additionally, the two Rosaceae-specific clades share a common origin with the CAD6/7/8 clade. Ka/Ks analysis was performed on gene pairs within each cluster. The median Ka/Ks values across the four clusters were generally similar; however, Group I exhibited higher extreme Ka/Ks values than the other three groups, suggesting that some genes in Group I have undergone strong positive selection during evolution (Fig. 4d).

COMT gene duplication and functional divergence correlate with lignification

A total of 450 *COMT* gene family members were identified from the 37 Rosaceae species and subjected to phylogenetic analysis. The results showed that the *COMT* members in Rosaceae clustered into four major clades (Fig. 5a). Analysis of motif frequencies across these four clades revealed generally consistent distributions for most motifs. However, the frequencies of motifs 7, 10, 11, 12, 13, and 15 were significantly lower in Clade I compared to the other clades. Similarly, the frequency of motif 13 was reduced in Clade IV, suggesting potential functional divergence of members in these two clades, particularly in Clade I (Fig. 5b). Furthermore, a syntenic network of the *COMT* gene family was constructed in Rosaceae. The network revealed that the *COMT* genes can be divided into four gene clusters (Fig. 5c), indicating that the Rosaceae *COMT* members may have originated from four ancestral genes, which is consistent with the phylogenetic clustering. Ka/Ks analysis of gene pairs within the different clusters showed that Group I and Group IV exhibited higher Ka/Ks ratios compared to the other two groups, indicating that some genes in Group I and Group IV have

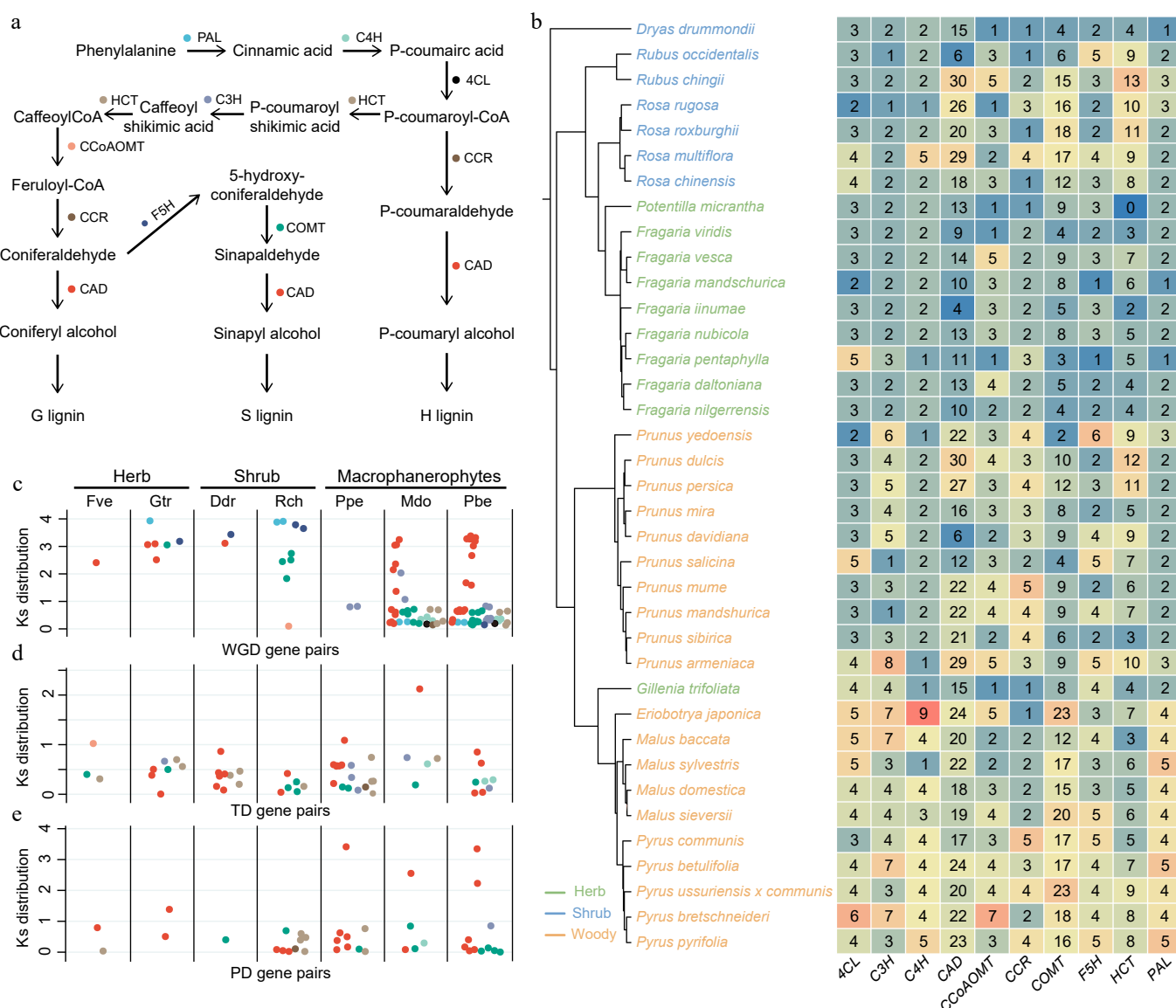


Fig. 3 Expansion of lignin biosynthesis genes in woody Rosaceae species is driven by recent duplication events. (a) Schematic diagram of the lignin biosynthesis pathway in plants. (b) Gene copy number variation in the lignin biosynthesis pathway across 37 Rosaceae species. The plants used in this study were all diploid. The color of labels represent different plant life forms. PAL, phenylalanine ammonia lyase; C4H, cinnamate-4-hydroxylase; 4CL, 4-coumarate:CoA ligase; HCT, hydroxycinnamoyl-CoA:shikimate/quinic acid hydroxycinnamoyltransferase; C3H, p-coumaroyl shikimate 3'-hydroxylase/coumaroyl 3-hydroxylase; CCoAOMT, trans-caffeoyl-CoA 3-O-methyltransferase; CCR, cinnamoyl-CoA reductase; F5H, ferulate 5-hydroxylase; COMT, caffeic acid 3-O-methyltransferase; CAD, cinnamyl alcohol dehydrogenase. C, D, E: Comparison of (c) whole-genome duplication, (d) tandem duplication, and (e) proximal duplication events for genes in the lignin biosynthesis pathway. The color of each point corresponds to the gene color in (a).

undergone positive selection during evolution, which may be related to the observed motif evolution (Fig. 5d).

A woody-specific HCT clade originated from gene duplication and was retained in lignified species

A total of 318 HCT gene family members were identified from the 37 plant species. Phylogenetic analysis revealed that the Rosaceae HCT genes are divided into three clades, with Clade II being specifically unique to woody Rosaceae species (Fig. 6a). Analysis of motif frequencies among the three clades showed that Clade I had the lowest motif frequency, while Clade III exhibited the highest (Fig. 6b), indicating that the gene sequences in Clade III are the most conserved, whereas those in Clade I are the least conserved. The syntenic network of HCT genes demonstrated that the three clades are grouped into two distinct clusters, with Clade I and Clade II originating from a common

ancestral source (Fig. 6c). Furthermore, Clade II, which is specifically present in woody Rosaceae species, likely originated from a duplication event of Clade I. Comparison of Ka/Ks values between Group I and Group II showed that Group I has a higher average Ka/Ks value than Group II, which suggests that Group I may have been under stronger selective pressure during evolution.

Discussion

Lignin, as a key component of the plant secondary cell wall, plays a central role in mechanical support, water transport, and stress response^[37]. This study has, for the first time, systematically investigated the evolutionary history of key gene families in the lignin biosynthesis pathway within Rosaceae, a family characterized by remarkable life form diversity. The findings provide new insights into

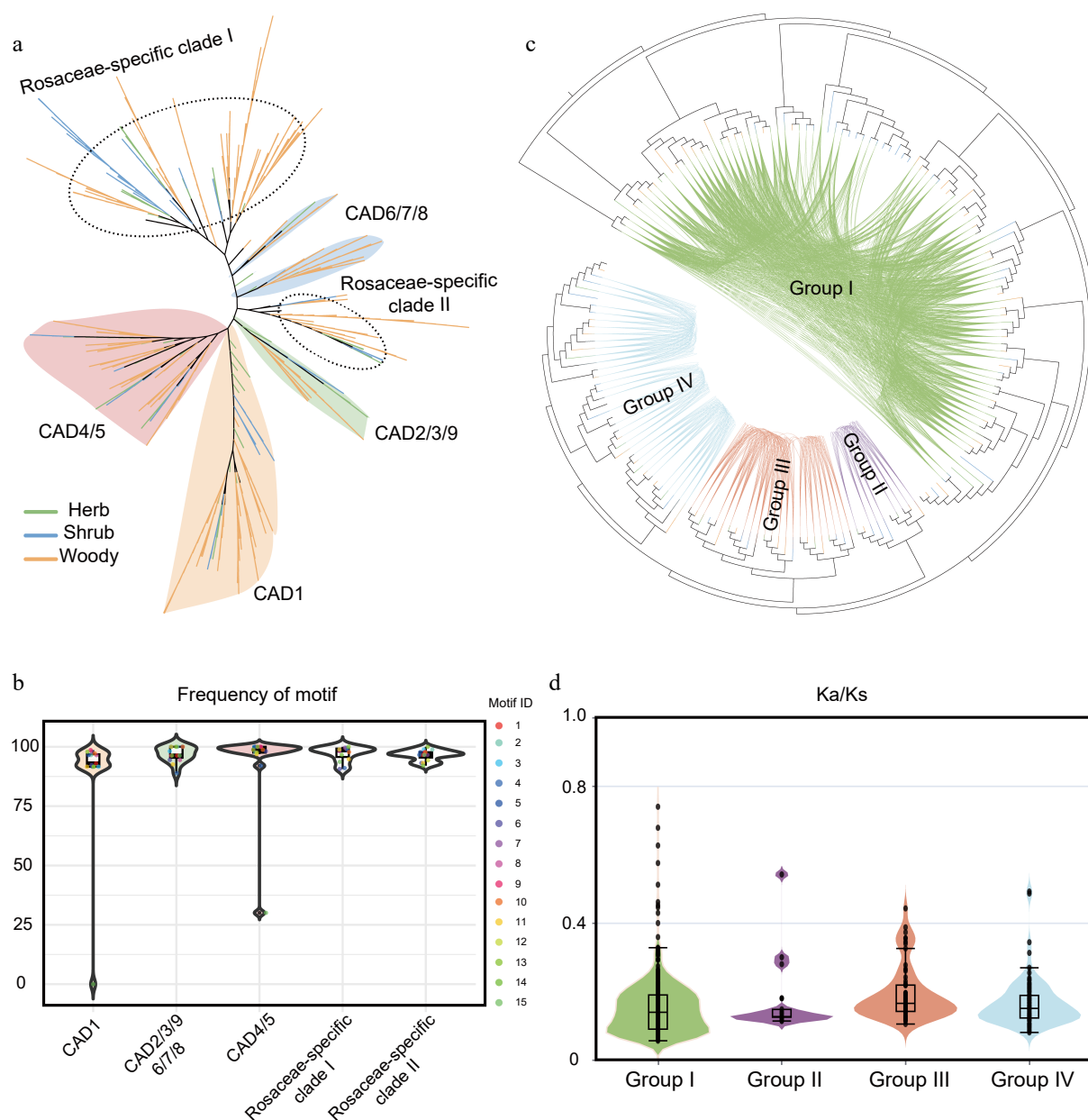


Fig. 4 Phylogenetic history and selection pressure of the *CAD* gene family in Rosaceae. (a) Phylogenetic analysis of *CAD* gene family members from 37 Rosaceae species. Distinct background colors represent different clades, while branch colors indicate distinct life-form plant groups. (b) Motif frequency distribution across different evolutionary clades. Background colors indicate phylogenetic clades. (c) Syntenic network analysis of *CAD* members in Rosaceae. (d) Ka/Ks analysis of gene pairs within the four gene clusters. Background colors correspond to the group colors in (c).

the genetic mechanisms underlying the divergence of herbaceous, shrub, and tree life forms.

This study demonstrates that the evolution of life forms in Rosaceae, particularly the acquisition of woody traits, is closely linked to the expansion of gene families in the lignin biosynthesis pathway and specific gene duplication events. Phylogenomic analyses revealed that at key nodes (Nodes A and D) within woody lineages (such as apple and pear in *Amygdaloideae*), significant expansions occurred in gene families related to cellulose synthesis, lignin biosynthesis, and cell wall biogenesis. In contrast, these gene families predominantly exhibited a contraction trend in herbaceous lineages. In bamboo evolution, whole-genome duplication events have led to the expansion of gene families involved in cellulose and lignin biosynthesis, promoting the emergence of bamboo varieties with high lignification and enhanced mechanical resistance, which

aligns with the findings of this study^[38]. This pattern clearly indicates that enhanced lignin biosynthesis capacity represents a key genetic basis for the evolution of woody life forms in Rosaceae. This finding is consistent with previous studies demonstrating that whole-genome duplication serves as a major driving force in plant evolution, influencing multiple plant traits^[39,40]. Further in-depth analysis of three key gene families, *CAD*, *COMT*, and *HCT*, revealed that the increase in gene copy numbers in woody plants primarily originated from whole-genome duplication (WGD), tandem duplication (TD), and proximal duplication (PD) events. Particularly noteworthy, in typical woody plants such as apple and pear, we detected gene pairs derived from recent WGD events (with *Ks* values between 0–1), which provided them with a substantial number of additional gene copies. These copies may have further optimized lignin synthesis and deposition through functional divergence or expression

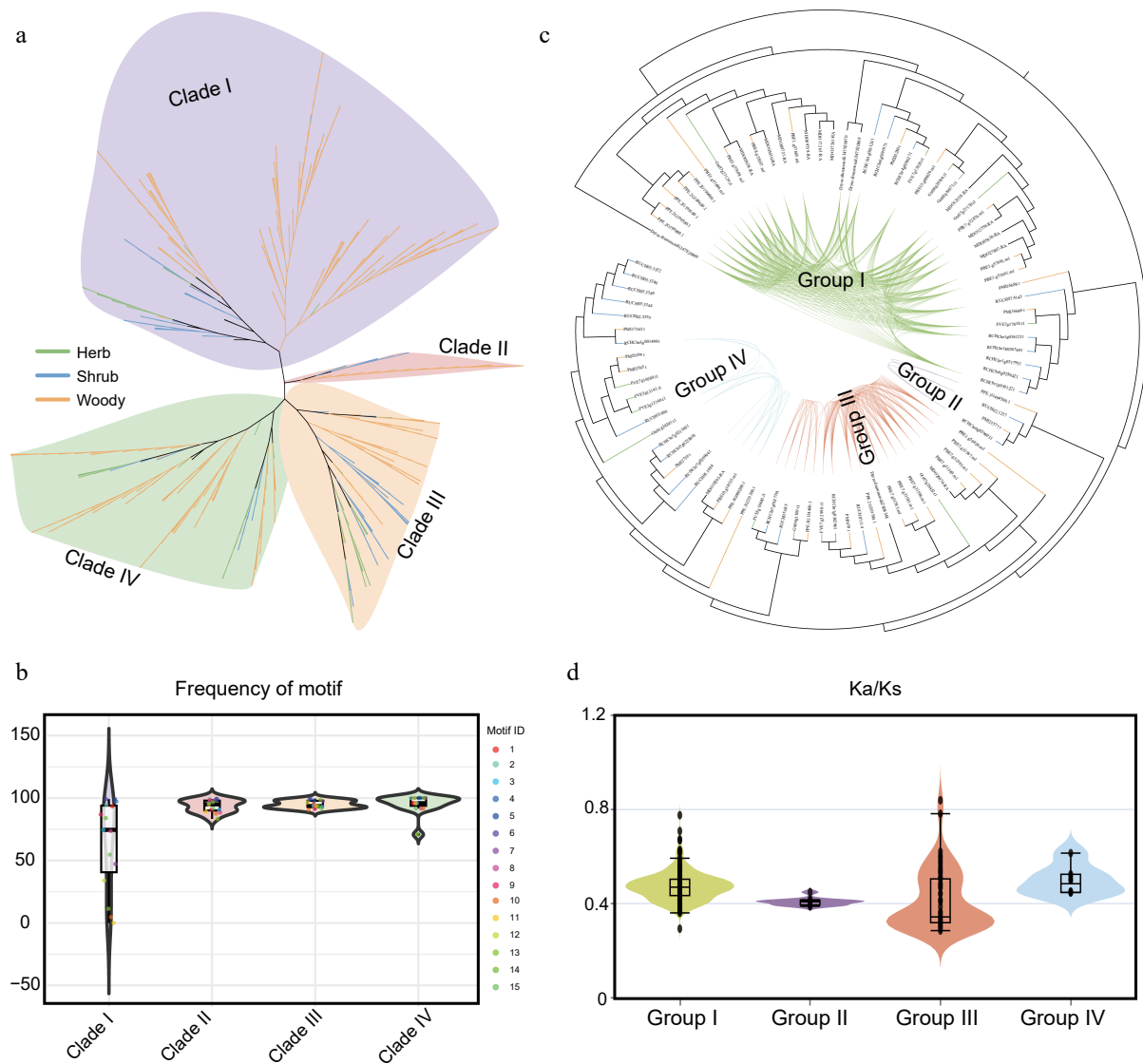


Fig. 5 Structural and evolutionary dynamics of the *COMT* gene family in Rosaceae. (a) Phylogenetic analysis of *COMT* gene family members from 37 Rosaceae species. Distinct background colors represent different clades, while branch colors indicate distinct life-form plant groups. (b) Motif frequency distribution across different evolutionary clades. Background colors indicate phylogenetic clades. (c) Syntenic network analysis of *COMT* members in Rosaceae. (d) Ka/Ks analysis of gene pairs within the four gene clusters. Background colors correspond to the group colors in (c).

specialization, thereby supporting their arborescent traits. Whole-genome duplication events and various types of gene duplications serve as major mechanisms for gene family expansion. Throughout evolution, genes that are retained following expansion and loss events are those preserved because they confer adaptive advantages for plant growth and development^[41,42].

This study identified *CAD*, *COMT*, and *HCT* as the three most frequently duplicated gene families in the lignin biosynthesis pathway of Rosaceae, whereas previous research on these gene families has been confined to single-species analyses, offering limited insights into their contribution to plant life forms^[43–45]. *CAD*, involved in the final crucial step of the plant lignin biosynthesis pathway, exhibits a multi-gene characteristic, suggesting that its members may have functional differentiation or redundancy across different tissues, developmental stages, and environmental conditions^[46]. In rice and sorghum, *CAD* members can be classified into several functional subgroups, with different *CAD* genes showing differential expression based on tissue (root, stem, leaf) or stress conditions (drought, high salinity)^[47,48]. Furthermore, studies have indicated

that *CAD* also plays significant roles in plant biotic stress responses^[48,49]. The elevated Ka/Ks ratio observed in Group I in this study (Fig. 4d) suggests that certain members have undergone positive selection, indicating that *CAD* genes in Group I might have acquired new or enhanced functions through adaptive changes in specific Rosaceae species. For instance, responses to environmental pressures such as low temperature, drought, or pathogen infection could be the primary driving factors. The functions of *COMT* in plants are also diverse, playing important roles in secondary metabolism and stress responses^[50,51]. The connection to functional differentiation in lignin synthesis and secondary metabolism implies complex gene duplication events during evolution^[52]. The distinct motif patterns and higher evolutionary rates observed in Clades I and IV (Fig. 5b, d) suggest that these duplicated clades may have separated ancestral functions, potentially refining the spatiotemporal control of S-lignin biosynthesis. The *HCT* gene family plays key roles in secondary metabolism, lignin synthesis, and environmental adaptation^[53–55]. Previous studies have observed that the *HCT* gene family is widely distributed across various plant lineages^[56] and has

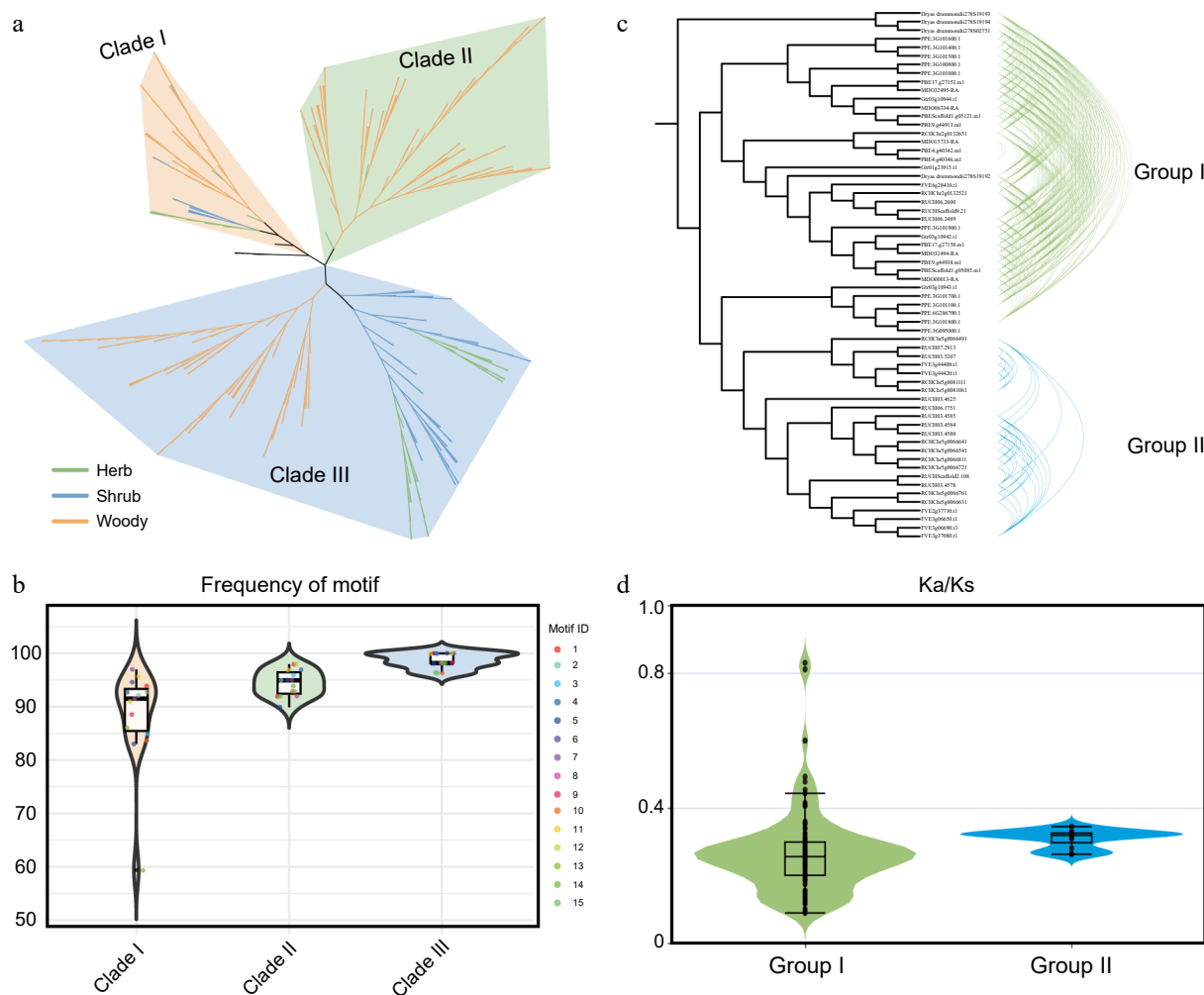


Fig. 6 Emergence of a woody-specific *HCT* clade through gene duplication. (a) Phylogenetic analysis of *HCT* gene family members from 37 Rosaceae species. Distinct background colors represent different clades, while branch colors indicate distinct life-form plant groups. (b) Motif frequency distribution across different evolutionary clades. Background colors indicate phylogenetic clades. (c) Syntenic network analysis of *HCT* members in Rosaceae. (d) Ka/Ks analysis of gene pairs within the two gene clusters. Background colors correspond to the group colors in (c).

expanded in land plants through genome duplication^[57]. The emergence of Clade II observed in this study, which is retained only in woody Rosaceae species and originated from a duplication of Clade I (Fig. 6a, c), may represent a key molecular adaptation in woody Rosaceae plants.

In summary, this study, through comparative genomic analysis of 37 Rosaceae species, reveals that the expansion and contraction of gene families in the lignin biosynthesis pathway, along with the duplication and functional differentiation of key genes (*CAD*, *COMT*, *HCT*), serve as important molecular driving forces behind the diversification of life forms from herbs to shrubs and trees within the family. These findings not only deepen our understanding of the genetic basis underlying the evolution of plant mechanical traits but also provide valuable genetic resources and a theoretical foundation for molecular breeding in Rosaceae.

Author contributions

The authors confirm their contributions to the paper as follows: experiments conception and design: Su LY and Xiong AS; data analysis and manuscript writing: Su LY, Liu ZT, Chen PY, Wang XL, and Xiong JS. All authors read and approved the final manuscript.

Date availability

All data generated or analyzed during this study are included in this published article.

Acknowledgments

The research was supported by the National Natural Science Foundation of China (32502666), Natural Science Foundation of Jiangsu Province (BK20251535), Postdoctoral Innovation Program of Shandong Province (SDCX-ZG-202503170), Fundamental Research Funds for the Central Universities (KYLH2025002), Priority Academic Program Development of Jiangsu Higher Education Institutions Project (PAPD), and the Bioinformatics Center of Nanjing Agricultural University.

Conflict of interest

The authors declare that they have no conflict of interest.

Supplementary information accompanies this paper at (<https://www.maxapress.com/article/doi/10.48130/gcomm-0025-0025>)

Dates

Received 11 October 2025; Revised 27 November 2025; Accepted 4 December 2025; Published online 29 December 2025

References

- Deng T, Du Q, Zhu Y, Queenborough SA. 2025. Environmental drivers of herbaceous plant diversity in the understory community of a warm-temperate forest. *Plant Diversity* 47:282–90
- Folk RA, Siniscalchi CM, Soltis DE. 2020. Angiosperms at the edge: Extremity, diversity, and phylogeny. *Plant, Cell & Environment* 43:2871–93
- Qian H, Jin Y, Ricklefs RE. 2017. Phylogenetic diversity anomaly in angiosperms between eastern Asia and eastern North America. *Proceedings of the National Academy of Sciences of the United States of America* 114:11452–57
- Murphy SJ, Salpeter K, Comita LS. 2016. Higher β -diversity observed for herbs over woody plants is driven by stronger habitat filtering in a tropical understory. *Ecology* 97:2074–84
- Xiang Y, Huang CH, Hu Y, Wen J, Li S, et al. 2017. Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Molecular Biology and Evolution* 34:262–81
- Vanholme R, Demedts B, Morreel K, Ralph J, Boerjan W. 2010. Lignin biosynthesis and structure. *Plant Physiology* 153:895–905
- Han X, Zhao Y, Chen Y, Xu J, Jiang C, et al. 2022. Lignin biosynthesis and accumulation in response to abiotic stresses in woody plants. *Forestry Research* 2:9
- Chen K, Guo Y, Song M, Liu L, Xue H, et al. 2020. Dual role of MdSND1 in the biosynthesis of lignin and in signal transduction in response to salt and osmotic stress in apple. *Horticulture Research* 7:204
- Tu M, Wang X, Yin W, Wang Y, Li Y, et al. 2020. Grapevine *VbZIP30* improves drought resistance by directly activating *VvNAC17* and promoting lignin biosynthesis through the regulation of three peroxidase genes. *Horticulture Research* 7:150
- Zhao Q. 2016. Lignification: flexibility, biosynthesis and regulation. *Trends in Plant Science* 21:713–21
- Li YP, Su LY, Huang T, Liu H, Tan SS, et al. 2025. The telomere-to-telomere genome of Pucai (蒲菜) (*Typha angustifolia* L.): a distinctive semi-aquatic vegetable with lignin and chlorophyll as quality characteristics. *Horticulture Research* 12:uhaf079
- Tuskan GA, Muchero W, Tschaplinski TJ, Ragauskas AJ. 2019. Population-level approaches reveal novel aspects of lignin biosynthesis, content, composition and structure. *Current Opinion in Biotechnology* 56:250–57
- Xiao P, Pfaff SA, Zhao W, Debnath D, Vojvodin CS, et al. 2025. Emergence of lignin-carbohydrate interactions during plant stem maturation visualized by solid-state NMR. *Nature Communications* 16:8010
- Gusakova M, Khviyuzov S, Bogolitsyn K, et al. 2024. Changes in the content of the main components of wood during the life cycle of higher plants. *Proceedings of the National Academy of Sciences, India Section B: Biological Sciences* 94:727–32
- Wang Y, Gui C, Wu J, Gao X, Huang T, et al. 2022. Spatio-temporal modification of lignin biosynthesis in plants: a promising strategy for ligno-cellulose improvement and lignin valorization. *Frontiers in Bioengineering and Biotechnology* 10:917459
- Ali Shad M, Li X, Rao MJ, Luo Z, Li X, et al. 2024. Exploring lignin biosynthesis genes in rice: evolution, function, and expression. *International Journal of Molecular Sciences* 25:10001
- Zhan W, Cui L, Song N, Liu X, Guo G, et al. 2025. Comprehensive analysis of cinnamoyl-CoA reductase (*CCR*) gene family in wheat: implications for lignin biosynthesis and stress responses. *BMC Plant Biology* 25:567
- Lin SJ, Yang YZ, Teng RM, Liu H, Li H, et al. 2021. Identification and expression analysis of caffeoyl-coenzyme A O-methyltransferase family genes related to lignin biosynthesis in tea plant (*Camellia sinensis*). *Protoplasma* 258:115–27
- Yin T, Xu R, Zhu L, Yang X, Zhang M, et al. 2024. Comparative analysis of the PAL gene family in nine citrus species provides new insights into the stress resistance mechanism of *Citrus* species. *BMC Genomics* 25:1020
- Wu P, Zhang R, Yu S, Fu J, Guo Z, et al. 2023. Genome-wide identification and expression analysis of the CAD gene family in walnut (*Juglans regia* L.). *Biochemical Genetics* 61:1065–85
- Weng JK, Chapple C. 2010. The origin and evolution of lignin biosynthesis. *New Phytologist* 187:273–85
- Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* 20:238
- Mendes FK, Vanderpool D, Fulton B, Hahn MW. 2021. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* 36:5516–18
- Wu T, Hu E, Xu S, Chen M, Guo P, et al. 2021. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *The Innovation* 2:100141
- Altschul S. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215:403–10
- Blum M, Andreeva A, Florentino LC, Chuguransky SR, Grego T, et al. 2025. InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Research* 53:D444–D456
- Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490
- Tang H, Krishnakumar V, Zeng X, Xu Z, Taranto A, et al. 2024. JCVI: a versatile toolkit for comparative genomics analysis. *iMeta* 3:e211
- Qiao X, Li Q, Yin H, Qi K, Li L, et al. 2019. Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. *Genome Biology* 20:38
- Zhang Z, Xiao J, Wu J, Zhang H, Liu G, et al. 2012. ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. *Biochemical and Biophysical Research Communications* 419:779–81
- Zhang Z. 2022. KaKs_Calculator 3.0: calculating selective pressure on coding and non-coding sequences. *Genomics, Proteomics & Bioinformatics* 20:536–40
- Rozewicki J, Li S, Amada KM, Standley DM, Katoh K. 2019. MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Research* 47:W5–W10
- Smith SA, Dunn CW. 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* 24:715–16
- Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27:1164–65
- Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, et al. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution* 37:1530–34
- Bailey TL, Boden M, Buske FA, Frith M, Grant CE, et al. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research* 37:W202–W208
- Vanholme R, De Meester B, Ralph J, Boerjan W. 2019. Lignin biosynthesis and its integration into metabolism. *Current Opinion in Biotechnology* 56:230–39
- Guo ZH, Ma PF, Yang GQ, Hu JY, Liu YL, et al. 2019. Genome sequences provide insights into the reticulate origin and unique traits of woody bamboos. *Molecular Plant* 12:1353–65
- Badouin H, Velt A, Gindraud F, Flutre T, Dumas V, et al. 2020. The wild grape genome sequence provides insights into the transition from dioecy to hermaphroditism during grape domestication. *Genome Biology* 21:223
- Droc G, Martin G, Guignon V, Summo M, Sempéré G, et al. 2022. The banana genome hub: a community database for genomics in the Musaceae. *Horticulture Research* 9:uhac221
- Conant GC, Birchler JA, Pires JC. 2014. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Current Opinion in Plant Biology* 19:91–98
- Liu S, Liu Y, Yang X, Tong C, Edwards D, et al. 2014. The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. *Nature Communications* 5:3930
- Zhang X, Wang Z, Zhong X, Fu W, Li Y, et al. 2024. Genome-wide identification of the *Gossypium hirsutum* CAD gene family and functional study of *GhiCAD23* under drought stress. *PeerJ* 12:e18439
- Liu Y, Wang Y, Pei J, Li Y, Sun H. 2021. Genome-wide identification and characterization of *COMT* gene family during the development of blueberry fruit. *BMC Plant Biology* 21:5

45. Ma C, Zhang H, Li J, Tao S, Qiao X, et al. 2017. Genome-wide analysis and characterization of molecular evolution of the *HCT* gene family in pear (*Pyrus bretschneideri*). *Plant Systematics and Evolution* 303:71–90
46. Peracchi LM, Brew-Appiah RAT, Garland-Campbell K, Roalson EH, Sanguinet KA. 2024. Genome-wide characterization and expression analysis of the cinnamyl alcohol dehydrogenase gene family in *Triticum aestivum*. *BMC Genomics* 25:816
47. Liu Q, Luo L, Wang X, Shen Z, Zheng L. 2017. Comprehensive analysis of rice laccase gene (*OsLAC*) family and ectopic expression of *OsLAC10* enhances tolerance to copper stress in *Arabidopsis*. *International Journal of Molecular Sciences* 18:209
48. Saballos A, Ejeta G, Sanchez E, Kang C, Vermerris W. 2009. A genomewide analysis of the cinnamyl alcohol dehydrogenase family in sorghum [*Sorghum bicolor* (L.) moench] identifies *SbCAD2* as the *brown midrib6* gene. *Genetics* 181:783–95
49. Chen C, Chang J, Wang S, Lu J, Liu Y, et al. 2021. Cloning, expression analysis and molecular marker development of cinnamyl alcohol dehydrogenase gene in common wheat. *Protoplasma* 258:881–89
50. Huang E, Tang J, Song S, Yan H, Yu X, et al. 2024. Caffeic acid O-methyltransferase from *Ligusticum chuanxiong* alleviates drought stress, and improves lignin and melatonin biosynthesis. *Frontiers in Plant Science* 15:1458296
51. Pham TH, Tian X, Zhao H, Li T, Lu L. 2024. Genome-wide characterization of *COMT* family and regulatory role of *CsCOMT19* in melatonin synthesis in *Camellia sinensis*. *BMC Plant Biology* 24:51
52. Liu Y, Bian Z, Jiang S, Wang X, Jiao L, et al. 2025. Comparative genomic analysis of *COMT* family genes in three *Vitis* Species reveals evolutionary relationships and functional divergence. *Plants* 14:2079
53. Eudes A, Pereira JH, Yogiswara S, Wang G, Teixeira Benites V, et al. 2016. Exploiting the substrate promiscuity of hydroxycinnamoyl-CoA: shikimate hydroxycinnamoyl transferase to reduce lignin. *Plant & Cell Physiology* 57:568–79
54. Chen Y, Yi N, Yao SB, Zhuang J, Fu Z, et al. 2021. CsHCT-mediated lignin synthesis pathway involved in the response of tea plants to biotic and abiotic stresses. *Journal of Agricultural and Food Chemistry* 69:10069–81
55. Rezagholi M, Rezapour Fard J, Darvishzadeh R. 2025. Selenium nanoparticles mitigates drought stress in *E. purpurea* by enhancing morpho-physiological characteristics and gene expression related to the phenylpropanoid pathway. *Industrial Crops and Products* 227:120833
56. Kriegshauser L, Knosp S, Grienberger E, Tatsumi K, Gütle DD, et al. 2021. Function of the hydroxycinnamoyl-CoA: shikimate hydroxycinnamoyl transferase is evolutionarily conserved in embryophytes. *The Plant Cell* 33:1472–91
57. Zheng K, Cai Y, Qu Y, Teng L, Wang C, et al. 2024. Effect of the HCT gene on lignin synthesis and fiber development in *Gossypium barbadense*. *Plant Science* 338:111914



Copyright: © 2025 by the author(s). Published by Maximum Academic Press, Fayetteville, GA. This article is an open access article distributed under Creative Commons Attribution License (CC BY 4.0), visit <https://creativecommons.org/licenses/by/4.0/>.