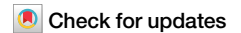


<https://doi.org/10.1038/s42003-025-07525-x>

Genome assembly of *Stewartia sinensis* reveals origin and evolution of orphan genes in Theaceae



Lin Cheng^{1,2,11}, Qunwei Han^{1,2,11}, Yanlin Hao¹, Zhen Qiao¹, Mengge Li¹, Daliang Liu^{3,4}, Hao Yin^{3,4}, Tao Li^{3,4}, Wen Long⁵, Shanshan Luo^{3,4}, Ya Gao^{3,4}, Zhihan Zhang^{3,4}, Houlin Yu^{6,10}, Xinhao Sun⁷, Hao Li^{8,9}✉ & Yiyong Zhao^{3,4}✉

Orphan genes play crucial roles in diverse biological processes, but the evolutionary trajectories and functional divergence remain largely unexplored. The Theaceae family, including the economically and culturally important tea plant, offers a distinctive model to examine these aspects. Here, we integrated Nanopore long-read sequencing, Illumina short-read sequencing, and Hi-C methods to decode a pseudo-chromosomal genome assembly of *Stewartia sinensis*, from the earliest-diverging tribe of Theaceae, spanning 2.95 Gb. Comparative genomic analysis revealed the absence of recent whole-genome duplication events in the Theaceae ancestor, highlighting tandem duplications as the predominant mechanism of gene expansion. We identified 31,331 orphan genes, some of which appear to have ancient origins, suggesting early emergence with frequent gains and losses, while others seem more specific and recent. Notably, orphan genes are distinguished by shorter lengths, fewer exons and functional domains compared to genes that originate much earlier, like transcription factors. Moreover, tandem duplication contributes significantly to the adaptive evolution and characteristic diversity of Theaceae, and it is also a major mechanism driving the origination of orphan genes. This study illuminates the evolutionary dynamics of orphan genes, providing a valuable resource for understanding the origin and evolution of tea plant flavor and enhancing genetic breeding efforts.

The family Theaceae, classified within the order Ericales, exhibits a remarkable level of biodiversity among angiosperms. With approximately 370 accepted species, the Theaceae family includes many economically and horticulturally important species¹, such as the tea plant (*Camellia sinensis*), *Camellia oleifera*, *Camellia japonica*, *Camellia sasanqua*. Recent studies incorporating morphological and molecular data have delineated three principal tribes within Theaceae: Theaeae, Gordoniaeae, and Stewartiaeae. The Stewartiaeae tribe, which is recognized as the sister to the remaining Theaceae species, encompasses the genera *Stewartia* and *Hartia*. Advanced phylogenetic analyses have provided insights into the evolutionary history

of these tribes, suggesting that the most recent common ancestor (MRCA) of the Stewartiaeae tribe likely originated approximately 20.78 million years ago^{2–4}. Previous morphological studies that considered *Stewartia gemmata* to be a synonym of *Stewartia sinensis*^{5,6}. Based on previous study by Lin et al. (2021), integrative approaches in phylogeny, ecology, and morphology were applied to the *Stewartia sinensis*. They proposed with two distinct species with the valid names of *Stewartia gemmata* for southern lineage and *Stewartia sinensis* for northern lineage⁷. *S. gemmata* is predominantly found in south China, including Hunan, Jiangxi, Fujian, Guangdong, and Yunnan, flourishing in mixed forests at altitudes of 900 to 1500 meters^{7–9}.

¹Dabie Mountain Laboratory, College of Tea and Food Science, Xinyang Normal University, Xinyang, China. ²Henan International Joint Laboratory of Tea-oil Tree Biology and High-Value Utilization, College of Tea and Food Science, Xinyang Normal University, Xinyang, China. ³Guizhou Key Laboratory of Functional Agriculture, College of Agriculture, Guizhou University, Guiyang, China. ⁴State Key Laboratory of Public Big Data, College of Computer Science and Technology, Guizhou University, Guiyang, China. ⁵Xinyang Normal University Library, Xinyang Normal University, Xinyang, China. ⁶Department of Biochemistry and Molecular Biology, University of Massachusetts Amherst, Amherst, USA. ⁷College of Science, Northeastern University, Boston, USA. ⁸School of Life Sciences, East China Normal University, Shanghai, China. ⁹Shanghai Institute of Eco-Chongming (SIEC), Shanghai, China. ¹⁰Present address: Broad Institute of MIT and Harvard, Cambridge, USA. ¹¹These authors contributed equally: Lin Cheng, Qunwei Han. ✉e-mail: lihao@bio.ecnu.edu.cn; yiyongzhao1991@gmail.com

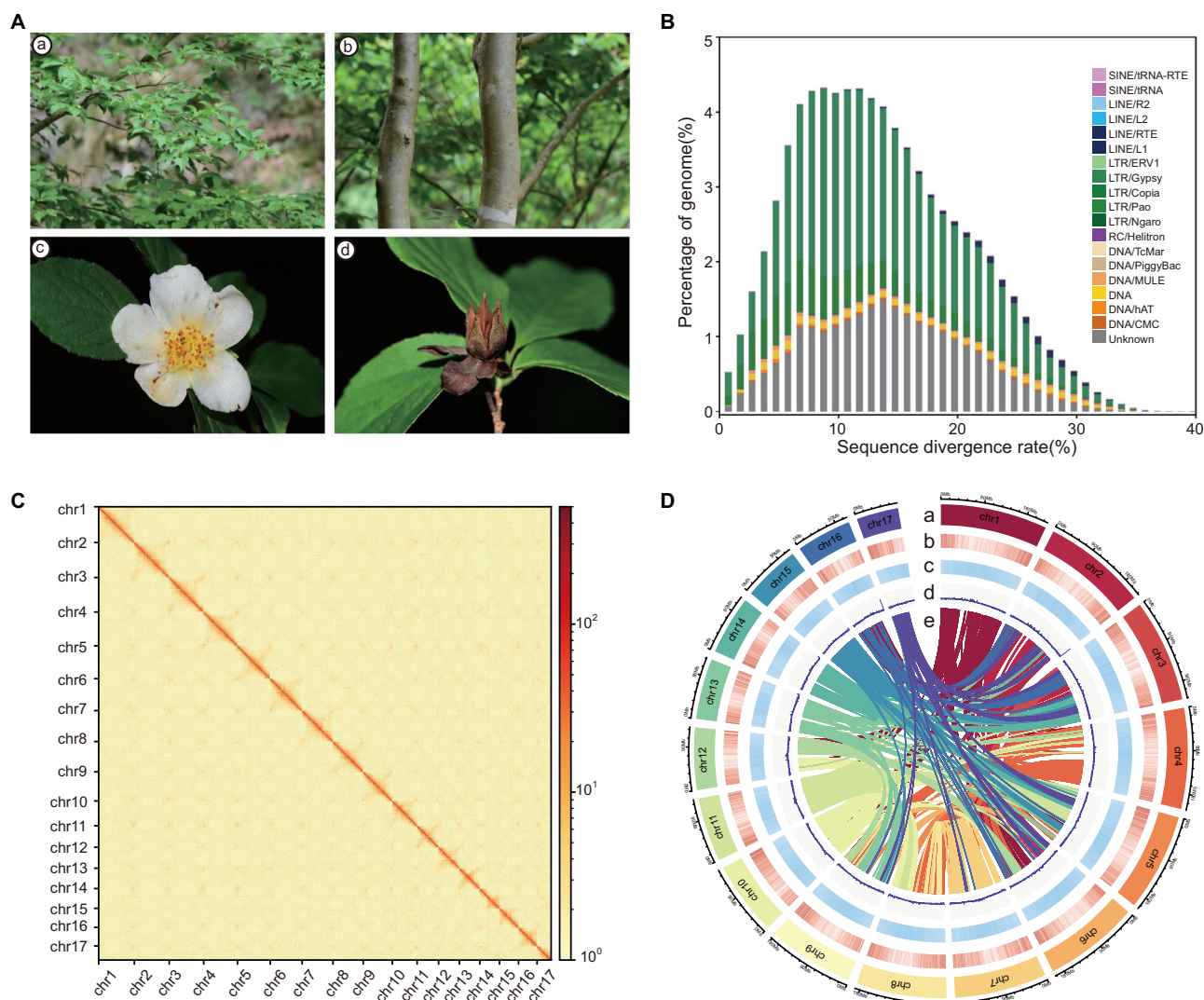


Fig. 1 | Morphological and genomic characteristics of *S. sinensis*.

A Morphological characteristics: Displaying various parts of *S. sinensis*. a-d illustrate the leaf, stem, flower and fruit, respectively, showcasing the distinctive morphology of *S. sinensis*. **B** Transposon sequence divergence distribution: The graph depicts the divergence between transposon sequences in the *S. sinensis* genome and their corresponding sequences in the Repbase database. The x-axis represents the divergence rate, while the y-axis shows the percentage of transposon sequences in the genome at each divergence level. Different transposon types are distinguished by unique colors. **C** Hi-C assisted genome assembly: Illustration of the chromosomal

structure of *S. sinensis* as determined by Hi-C assisted genome assembly. The horizontal and vertical axes represent different chromosomes. The color scale on the right indicates contact density, ranging from red (representing high density) to yellow (representing low density). **D** Genomic annotation information: This section details various genomic features of *S. sinensis*, including a: Depiction of the 17 chromosomes. b: Visualization of gene density across the chromosomes. c: Distribution of repeated sequences. d: GC content variation. e: Collinear connections between different chromosomes.

S. sinensis typically reaches a height between four and eight meters, characterized by its smooth and greyish-yellow bark^{6,10}. It is frequently used for ornamental purposes in horticulture owing to its vibrant flowers with high decorative and aesthetic appeal (Fig. 1A). The bark, roots, and fruits of this plant, used in traditional medicine, have significant medico-economic value⁵.

Orphan genes, also known as new genes or lineage-specific genes (LSGs), originate within the genomes of specific subsets of species¹¹. Extensive LSGs were identified from diverse eukaryotic lineages, including plants^{12–14}, animals^{15–20} and fungi^{21–23}. As innovative genetic elements, orphan genes drive functional and phenotypic diversity and significantly influence the evolutionary processes of organisms^{24,25}. Studying orphan genes relies on a stable and reliable phylogenetic foundation, with low-copy nuclear genes being particularly effective in discerning phylogenetic relationships among angiosperms^{26–28}. Compared to morphological, anatomical, and cytological traits, as well as mitochondrial and chloroplast genomic markers, nuclear

genes exhibit greater variability and longer sequence lengths. These characteristics enable nuclear genes to yield more detailed and comprehensive phylogenetic information, essential for understanding evolutionary histories and species diversification^{29,30}. Advances in genome sequencing have greatly facilitated the study of low-copy nuclear genes in plant phylogenetics^{31–33}. The genomes of *Amborella trichopoda* and *Nymphaea colorata* are instrumental in deepening our understanding of evolutionary mechanisms in early flowering plants^{34,35}. In bamboo, genomic analysis has uncovered 1622 orphan genes, shedding light on their role in the rapid growth of plant and offering insights into the genetic dynamics of bamboo stalk expansion³⁶. However, research on orphan genes in the Theaceae family is limited by the scarcity of genomic data from another genus.

Whole genome duplication (WGD), a key evolutionary mechanism, has been documented to have taken place in the common ancestor of extant seed plants, angiosperms, and core eudicots^{37–39}. WGD events have been identified in the early evolution of several major plant families, including

Asteraceae, Brassicaceae, Fabaceae, Poaceae, and Rosaceae^{40–44}. Studies utilizing the genome of *Actinidia chinensis*, commonly known as kiwifruit, have suggested the occurrence of a WGD event, termed WGD- β , which was believed to be common to both *Actinidia* and *Camellia* species in Ericales^{45,46}. Recent advancements in genomic analysis have illuminated the genomic localization of the WGD- β , predominantly associated with core Ericales, Primuloids, Polemonioids, and Lecythidaceae⁴⁷. Within the Theaceae family, genomic research primarily focuses on the genus *Camellia*, especially on *Camellia sinensis* (commonly known as tea plants) and Oil-Camellias species^{48–52}. Recent investigations into WGD analysis in the Theaceae family have revealed discrepancies, particularly concerning the number of WGD events post whole-genome triplication- γ (WGT- γ) and the possibility of an independent WGD event within the family^{39,46,49,53,54}.

Tea, a globally favored beverage, is produced from the leaves of *Camellia sinensis* (L.) O. Kuntze, a member of the Theaceae family⁵³. Tea consumption is linked to various health benefits, such as preventing low-density lipoprotein oxidation, reduction of serum cholesterol, and decreased risk of cardiovascular syndromes⁵⁵. Bioactive compounds in tea, including polyphenols, theanine and caffeine, are contributing to both flavor and potential health benefits⁵⁶. The complex nature of tea flavor is influenced by environmental conditions, cultivation practices, processing methods, and the selection of specific tea cultivars^{57–59}. WGD events, characterized by the complete duplication of an organism's genome, result in the introduction of additional copies of genes. This process significantly expands the pool of genetic material. Empirical evidence from prior research has demonstrated that WGD events contribute to increased genetic diversity within the tea plant (*Camellia sinensis*)^{46,49}. This increase in genetic diversity could potentially enhance the variation in alleles and genotypic compositions in tea plants⁶⁰. Moreover, WGD can impact gene expression in metabolic pathways⁶¹, including those responsible for the synthesis of flavor-related compounds. Changes in gene dosage resulting from WGD can lead to alterations in the production or accumulation of specific flavor compounds in tea leaves. While the WGD may contribute to the genetic foundation of tea flavor by generating novel genes through the neo-functionalization of redundant duplicated genes, a comprehensive understanding of these relationships and underlying mechanisms requires further investigation^{62–67}.

In our study, we successfully present the first pseudo-chromosomal genome assembly of *S. sinensis*, a species belonging to Stewartiae in the Theaceae family, using whole-genome sequencing of short and long-read sequencing technologies. We identified 31,331 orphan genes and numerous gene families within Theaceae that have undergone significant gene gain and loss events. Comparative genomics analysis revealed the absence of WGD events in Theaceae ancestors, with tandem duplications (TDs) being the predominant mechanism driving gene expansion and contributing to the adaptive evolution and characteristic diversity of the Theaceae. These findings greatly enhance our understanding of gene dynamics within Theaceae, elucidating their impact on both biodiversity and the diversification of tea flavor. Moreover, the study of orphan genes has the potential to advance our understanding of plant evolution, enhance crop improvement strategies, enable biotechnological applications, and contribute to the conservation and sustainable use of plant biodiversity.

Results

Genome sequencing, assembly and annotation

The genome size of *S. sinensis* was estimated to be approximately 2722 Mb based on flow cytometry using maize or soybean as reference species, respectively (Supplementary Fig. 1 and Supplementary Table 1). Its genome heterozygosity was 1.23% based on K-mer depth distribution analysis (Supplementary Fig. 2 and Supplementary Table 2). A comprehensive approach was employed for sequencing the *S. sinensis* genome, integrating next-generation sequencing (NGS) by BGI, Oxford Nanopore long-read sequencing, and HiC-based chromatin mapping. A total of 302.53 Gb of high-quality reads were obtained (Supplementary Table 3), representing 100.67-fold coverage of the *S. sinensis* genome. The Oxford Nanopore

sequencing generated 155.03 Gb of raw data, culminating in 154.31 Gb of clean data after filtering (Supplementary Table 4). From these processes, 753 contigs with a combined length of 2,951.35 Mb were produced, with the longest measuring 41.33 Mb (Supplementary Table 5). Aligning NGS reads against the reference genome showed a high mapping rate and coverage of 99.73% and 97.58%, respectively (Supplementary Table 6). BUSCO assessment with the embryophyta_odb10 dataset successfully assembled 1573 of 1614 genes, denoting 97.40% completeness (Supplementary Table 7). Furthermore, sequencing via the NovaSeq 6000 platform yielded 619.20 Gb of raw data, leading to 304 scaffolds with an N50 value of 190.65 Mb (Supplementary Table 8). Hi-C-assisted assembly facilitated the construction of 17 chromosomes, covering roughly 96.53% of the Hi-C assembled genome length of 2,848,999,739 bp (Supplementary Table 9). This bioinformatics result from Hi-C anchoring is consistent with previous studies on *Stewartia*, where chromosome numbers were determined using traditional cytology to calibrate genome sizes with ploidy levels. These studies confirmed a base chromosome number of $1n = 1x = 17$ through traditional cytology^{68,69}. A notable pattern was observed in the analysis, characterized by a single peak value at a sequence divergence rate of around 10 (Fig. 1B, Supplementary Table 10). The total length of these 17 chromosomes was 2.95 Gb, accounting for 96.53% of the total length of the Hi-C-assembled genome, suggesting that the pseudo-chromosome level of *S. sinensis* was consistent with its previously reported chromosome number by using traditional cytology⁶⁸ (Fig. 1C, D). Gene density, repeat sequence density, GC content, and synteny analyses, depicted in Fig. 1D, revealed higher gene density at chromosome ends compared to the middle regions.

The analyses indicated that a significant portion of the genome, approximately 2.37 Gb, consists of repeated sequences, accounting for about 80.18% of the total genome length (Supplementary Table 10), and long terminal repeats (LTRs) represent 59.93% of the genome. DNA transposons and RNA transposons were found to occupy 72.45 Mb (approximately 2.45% of the genome) and 1806.50 Mb (around 61.21% of the genome), respectively (Fig. 1B, Supplementary Table 10). A sequence homology-based prediction approach based on five closely related species annotated an average of 131,970 genes. This combined approach resulted in identifying a comprehensive set of 69,599 protein-coding genes (Supplementary Table 11). In comparison to related species such as *Camellia sinensis* var. *sinensis* (CSS) 'Tieguanyin', *Camellia lanceoleosa*, *Camellia* DASZ, *Vaccinium darrowii*, and *Actinidia chinensis*, the genes of *S. sinensis* exhibited a higher average number of exons (4.51 vs. 2.59) and exon length (305.5 bp vs. 260.71 bp). However, the average gene length of *S. sinensis* (9418 bp) was significantly shorter than that of related species (21,286 bp) (Supplementary Table 11). An extensive annotation analysis across multiple public databases indicated that a significant majority, 92.99%, of the genes had functional annotations (Supplementary Table 12). A total of 4647 ncRNA genes, 1023 transfer RNAs (tRNAs), 679 ribosomal RNAs (rRNAs), and 2669 small nuclear RNAs (snRNAs) were annotated within the genome (Supplementary Table 13).

The resolution of Theaceae phylogeny through genome-scale phylogenomic analysis

Constructing a comprehensive phylogenetic framework is vital for understanding evolutionary dynamics, genetic variation intricacies, and genome architecture biology, especially in the context of WGD events across diverse genomes. To establish an extensive phylogenetic tree of Theaceae, we utilized datasets comprising 13 genomic sequences and 150 transcriptomic sequences from Theaceae, along with four genomic sequences as outgroups (Supplementary Data 1, 2). The 13 samples in our study, represent high-quality Theaceae genomes, including *Camellia sinensis* varietals (*C. sinensis* var. *sinensis* 'Shuchazao', *C. sinensis* var. *sinensis* 'Longjing', *C. sinensis* var. *sinensis* 'Tieguanyin', *C. sinensis* var. *sinensis* 'Huangdan', *C. sinensis* 'DuyunMaojian', *C. sinensis* var. *sinensis* 'Biyun', *C. sinensis* var. 'YingHong9', *C. sinensis* var. *assamica* 'Yunkang 10', *Camellia* DASZ) and other Theaceae species such as *Camellia oleifera* var. *Nanyongensis*

BUSCO assessment results

- Complete and single-copy BUSCOs
- Complete and duplicated BUSCOs
- Fragmented BUSCOs
- Missing BUSCOs

Taxonomy

- Theaceae
- Gordoniaceae
- Stewartiaceae

Posterior probability

- 0.9 ≤ PP ≤ 1
- 0.8 ≤ PP < 0.9
- 0.6 ≤ PP < 0.8
- 0.4 ≤ PP < 0.6

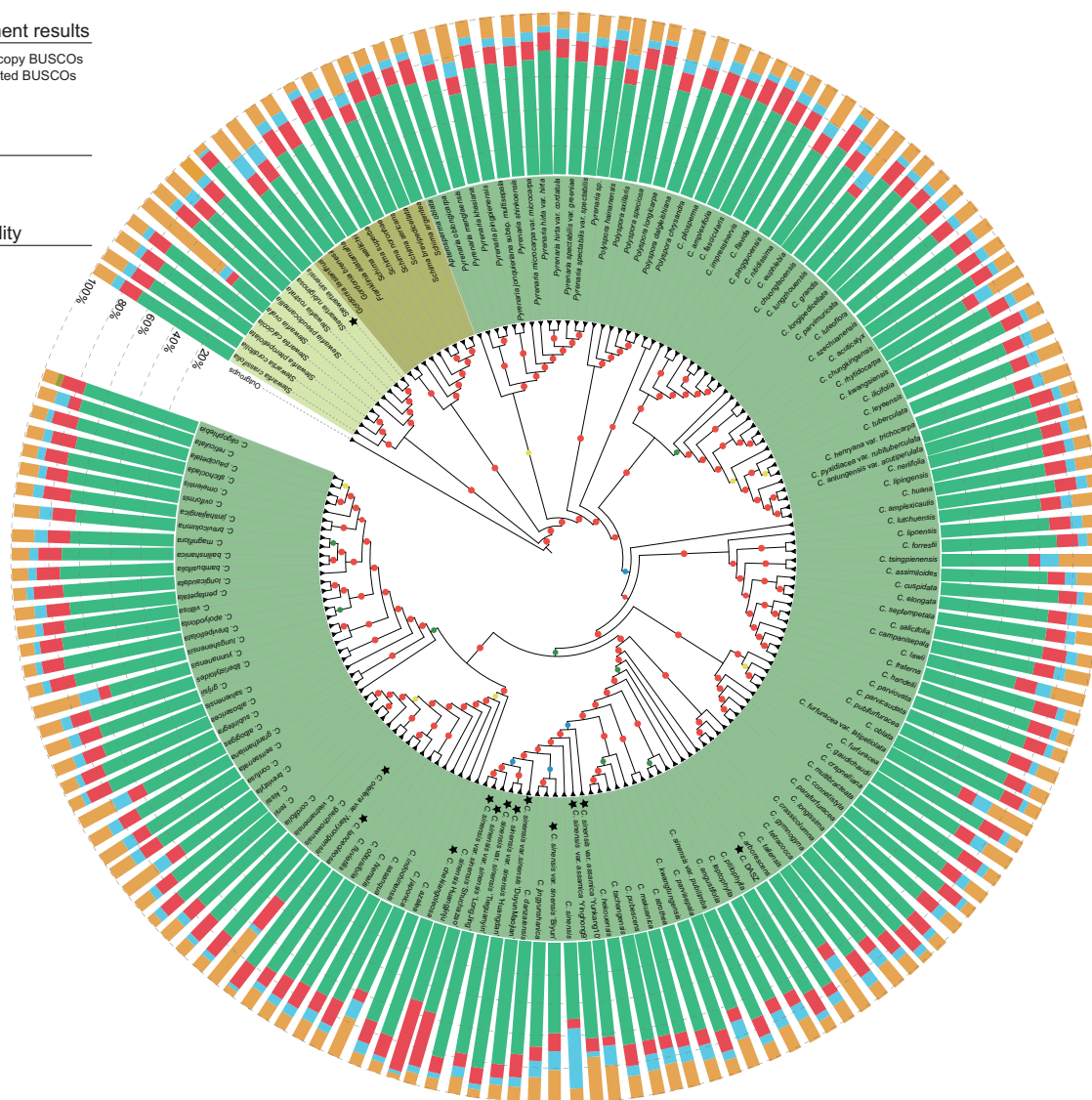


Fig. 2 | Phylogenetic relationship of Theaceae. The outer ring represents the BUSCO assembly results of the dataset, using the eudicots_odb10 reference database. The intermediate species names, marked by the letter C, signify the genus *Camellia*. The asterisk highlights the 13 genomic datasets within the Theaceae. The different colored dots in the inner circle indicating varying levels of support. A red

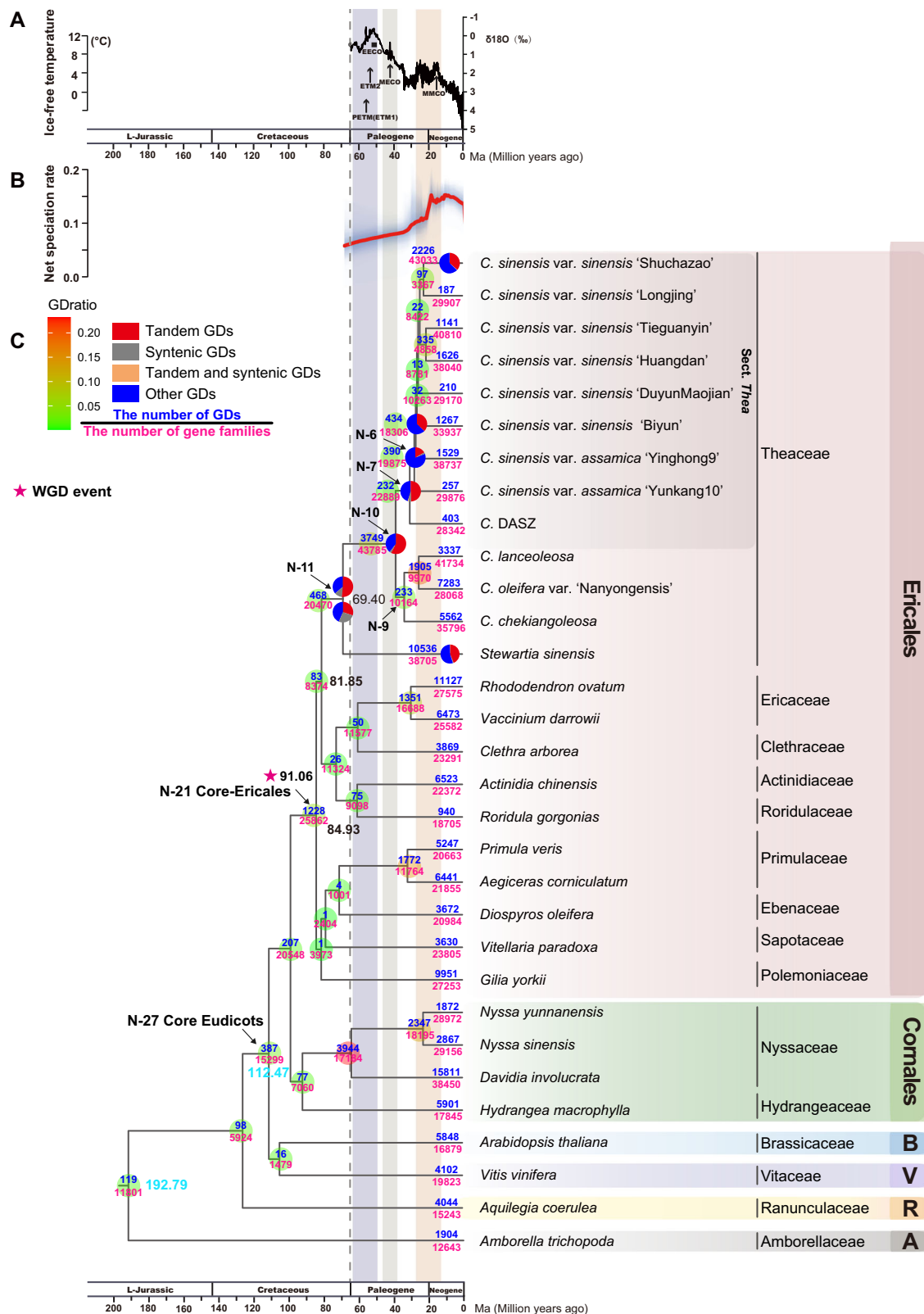
dot denotes a local posterior probability (PP) support value of 0.9 or higher, a yellow dot indicates a PP support value between 0.8 and 0.9, a green dot indicates a PP support value between 0.6 and 0.8 and a blue dot signifies a PP support value between 0.4 and 0.6.

(*C. oleifera* var. *Nanyongensis*) and *S. sinensis*, *C. lanceoleosa*, *C. chekiangoleosa*. These were thoroughly analyzed to evaluate genomic data completeness, facilitating in-depth phylogenomic analyses and orphan gene detection to elucidate evolutionary relationships and genomic diversity within the family. Notably, all genome/transcriptome assemblies displayed BUSCO (C) scores over 50% (Fig. 2) underscoring the high quality of the data used for the comprehensive phylogenetic analysis. The tribe Stewartiaceae emerged as sister groups to tribes Gordoniaceae and Theaceae with maximal support value. The highly resolved phylogenetic reconstruction within Gordoniaceae and Stewartiaceae aligns remarkably with existing phylogenetic studies. For further comparative genomic analyses, we integrated our findings with the established Asterids phylogenetic framework, which features extensive species coverage at the order level⁷⁰. In Ericales, Theaceae is positioned as the sister group to the most recent common ancestors of Roridulaceae, Actinidiaceae, Clethraceae, and Ericaceae families (Supplementary Fig. 3). Based on these results, we constructed a comprehensive species tree encompassing 31 genomes, which is detailed in Supplementary Fig. 3 with clearly labeled distinct nodes (Supplementary Data 2).

Gene family expansion and gene duplication in Theaceae

A previous study conducted an analysis of the global climate trend over the past 65 million years (Fig. 3A)^{27,71}. In our examination of the rate-through-time plot for Theaceae, we noted a significant increase in the speciation rate starting from the Core Tr. Theaceae during the late Oligocene, around 30.8 million years ago (Mya) (Fig. 3B)³. In order to examine the gene contents, potential expansions of gene families, and occurrences of large-scale gene duplication events throughout the evolutionary history of Theaceae, we conducted comparative genomic analysis on a dataset consisting of 13 genomes from Theaceae species. *S. sinensis* was chosen as the representative species from the early-diverging clade of Theaceae, while 18 additional angiosperm genomes were included as outgroups for comparison (Fig. 3C, Supplementary Table 14). Tricolpate pollen fossils were used as fossil calibration points for core eudicots plant stem node. The divergence time of Theaceae was estimated to be around 69.40 Mya, while *Camellia* diverged approximately 38.77 Mya (Supplementary Fig. 4).

Our study revealed numerous gene duplications (GDs) occurred at ancestral nodes of Theaceae. Specifically, at the node representing the most



recent common ancestor (MRCA) of core eudicots (N-27), we discovered 387 GDs. This finding echoes the well-documented whole-genome triplication- γ (WGT- γ) and subsequent WGD events within core Ericales (WGD- β), where we identified 1228 GDs at the Ericales node (N-21), as illustrated in Fig. 3C, Supplementary Table 14 and elaborated in Supplementary Fig. 3. Further investigations within the Theaceae family unveiled a

widespread occurrence of GDs across various nodes. At the MRCA of Theaceae (N-11), our analysis detected 468 GDs, representing 2.29% of the total gene families at that node. Remarkably, out of these duplications, only 130 gene duplicated pairs showcased a fully ABAB gene duplication pattern, comprising 27.78% of the gene pairs retained. A substantial number of duplications, 3749 in total, were observed at the MRCA of *Camellia* (N-10),

Fig. 3 | Divergence time of Ericales and identification of gene duplication events in Theaceae. **A** Global temperature curve and climatic events: This panel illustrates the global temperature curve and significant climatic events over the last 65 million years^{27,71}. EECO Early Eocene climatic optimum, MECO Mid-Eocene climatic optimum, MMCO Mid-Miocene climatic optimum, PETM (ETM1) Paleocene-Eocene thermal maximum 1, ETM2 Eocene thermal maximum 2. **B** Diversification rate in Ericales: The rate-through-time plot of Ericales is displayed here. The red line indicates the median diversification rate (species/million years), and the surrounding gray shadow represents the 95% credibility interval. **C** Detection of gene duplication: This section focuses on the detection of gene duplication events in Theaceae, based on the reconciliation of gene trees and species tree. The dashed gray line indicates the K-Pg boundary. The black numbers adjacent to the nodes

represent the age as predicted by the MCMCTree software. Pie charts displayed at each node depict the observed gene duplication types from CSS, while the pie chart specifically located beneath node N11 illustrates those identified in *S. sinensis*. Each color within the pie charts corresponds to a distinct type of gene duplication. Red: Proportion of gene duplications (GDs) from tandem duplications; Gray: Proportion of GDs resulting from synteny, indicating gene duplications that arise due to whole genome duplications event; Orange: Proportion of GDs where both tandem and synteny duplications occur simultaneously; Blue: Proportion of GDs attributed to other types of duplications not classified as tandem or synteny. The color of the circle at the node ranges from green to light red representing the ratio of GDs at the node to the total number of GDs ranging from small to large. B Brassicales, V Vitales, R Ranunculales, A Amborellales.

constituting 8.56% of the gene families present at that node. A predominant 80% of these duplications followed the ABAB pattern. This ABAB duplication trend was also present at the MRCA nodes of both cultivated tea (N-6) and the Oil-Camellias plant (N-9), with the GD ratios reflecting similar evolutionary patterns to those observed in other lineages within the genus *Camellia*.

In this study, intraspecific synteny analyses were conducted on *Camellia sinensis* var. *sinensis* 'Shuchazao' and *S. sinensis*. The genome of *S. sinensis* was found to comprise 2088 syntenic blocks distributed over 17 chromosomes. Dot plot analysis revealed prominent green syntenic blocks indicative of a 1:2 genomic correspondence in *S. sinensis* (Fig. 4A). This observation corroborates the previously reported whole-genome triplication event (WGT-γ) within the core eudicots. Additionally, orange syntenic blocks demonstrated a 1:1 correspondence, aligning with recent WGD events identified in the order Ericales (Fig. 4A). A similar synteny pattern was discerned in CSS 'Shuchazao', with 1909 syntenic blocks mapped to its 15 chromosomes (Supplementary Fig. 5A). These intraspecific synteny analysis suggest the absence of recent WGD events within the Theaceae family.

Furthermore, interspecific synteny analysis involving *S. sinensis* and other species, including CSS 'Shuchazao', *Vitellaria paradoxa*, and *Vitis vinifera*, revealed syntenic block correspondences of 1:1, 1:1, and 2:1, respectively (Supplementary Fig. 5B and Fig. 4B, C). The grape genome exhibited synteny with two chromosomes of *V. paradoxa*, denoting a 1:2 genomic correspondence. The synteny blocks of *S. sinensis*, CSS 'Shuchazao', and *V. paradoxa* displayed a 1:1 correspondence on the corresponding chromosomes (Fig. 4D), providing no evidence for WGD at the ancestral node of Theaceae. Post-divergence from *V. vinifera*, *S. sinensis* experienced an additional WGD event, shared with *V. paradoxa* and CSS 'Shuchazao' congruent with the WGD detected in the core Ericales lineage.

Analysis of synonymous substitution rates (Ks) in *S. sinensis*, CSS 'Shuchazao', *Vitis vinifera* and *V. paradoxa* revealed the evidence of core-Ericales WGD and WGT-γ events, with discernible peaks at Ks values approximating 0.5 and 1–1.5, respectively (Fig. 4E). The distribution of the Ks values about the orthologous gene pairs in comparing *S. sinensis* with CSS 'Shuchazao', *V. vinifera* and *V. paradoxa* (Fig. 4F) has further supported the hypothesis of an absence of a recent, lineage-specific WGD event in the common ancestor of Theaceae.

Identifications of gene duplication types in Theaceae

In the absence of WGD events explicitly identified within Theaceae, an in-depth investigation was undertaken to elucidate the nature of GD occurrences within this family. Our approach involved a meticulous gene tree-species tree reconciliation and an assessment of genomic synteny across various Theaceae taxa including CSS 'Shuchazao' and *S. sinensis*. The analysis revealed that a minority of GD pairs, which originated from ancestral Theaceae lineages at nodes N-11, N-10, N-7, and N-6, demonstrated synteny within their respective genomes. This observation suggests that the majority of these GD pairs were not generated through WGD events. Further scrutiny, quantifying the chromosomal proximity of these gene duplicates, indicated that gene pairs situated within a 10-gene radius comprised between 3.90% to 27.90% and 14.10% to 17% of the total duplicated

gene pairs in CSS 'Shuchazao' and *S. sinensis*, respectively, as depicted in Supplementary Fig. 6. In addition to the gene pairs of unknown origin, tandem duplication events underscore the significant role of the gene duplication landscape observed in Theaceae.

GO and KEGG enrichment analysis of tandem duplication genes

In order to explore the functional significance of tandem duplication in Theaceae, we conducted the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment for those genes under tandem duplications. During the ancestral node of Theaceae (N-11), genes associated with CSS 'Shuchazao' underwent tandem duplication and enriched in active metabolic pathways such as zein and catecholoxidase (Supplementary Fig. 7A). At the ancestral node of the *Camellia* genus (N-10), the genes with tandem duplication in CSS 'Shuchazao' were enriched in the defense responses, including oxidative stress response and biological stimuli, as well as flavonoid metabolism, terpene compounds, salicylic acid, and lignin synthesis (Supplementary Fig. 7A). Tandem duplication genes, in addition to their fundamental role in sustaining life, play a crucial role in enhancing adaptability to the environment and contributing to species-specific traits. CSS 'Shuchazao' and *S. sinensis* exhibit distinct responses in tandem duplication genes to defense mechanisms, with CSS 'Shuchazao' demonstrating responses to ultraviolet light and bacterial defense, while *S. sinensis* primarily shows fungal defense responses (Supplementary Fig. 7B). Furthermore, CSS 'Shuchazao' responds to the metabolism of terpenes, fatty acids, and flavonoids at various nodes, enriching the spectrum of responsive enzymes (Supplementary Fig. 7C). In contrast, *S. sinensis* associated with the metabolism of terpenoids, vitamins, lignin, xylem and phloem development (Supplementary Fig. 7D).

Gene family expansion in Theaceae ancestors

In the evolutionary history of angiosperms, the gain and loss of gene families have been pivotal, shaping species-specific characteristics. Through PhyloMCL⁷² clustering of 1,245,359 protein sequences from 31 angiosperm genomes, we identified a total of 195,197 homologous gene clusters. These gene clusters were mapped to the species tree to analyze gene family expansion and contraction dynamics during the evolution of Theaceae (Supplementary Fig. 8). This analysis indicated that 2350 gene families were gained at the MRCA of Theaceae. At the MRCA of the *Camellia* genus, a significant expansion to 7306 gene families was observed (Supplementary Fig. 8). Additionally, 2396 gene families were expanded at MRCA of cultivated teas, while 299 gene families were expanded at the MRCA of CSS (Supplementary Fig. 8). Moreover, at the MRCA of Oil-Camellias including *Camellia oleifera*, *C. lanceoleosa* and *C. chekiangoleosa* with a total of 2859 gene families were expanded. Gene family expansions provide a pivotal insight into the genetic underpinnings of Theaceae, particularly in the context of environmental adaptation and the development of species-specific characteristics.

Identification and distribution of orphan genes and transcription factors (TFs) of Theaceae

In the Theaceae family, comprehensive genomic analysis led to the identification of 31,331 orphan genes, as detailed in Supplementary Table 15 and

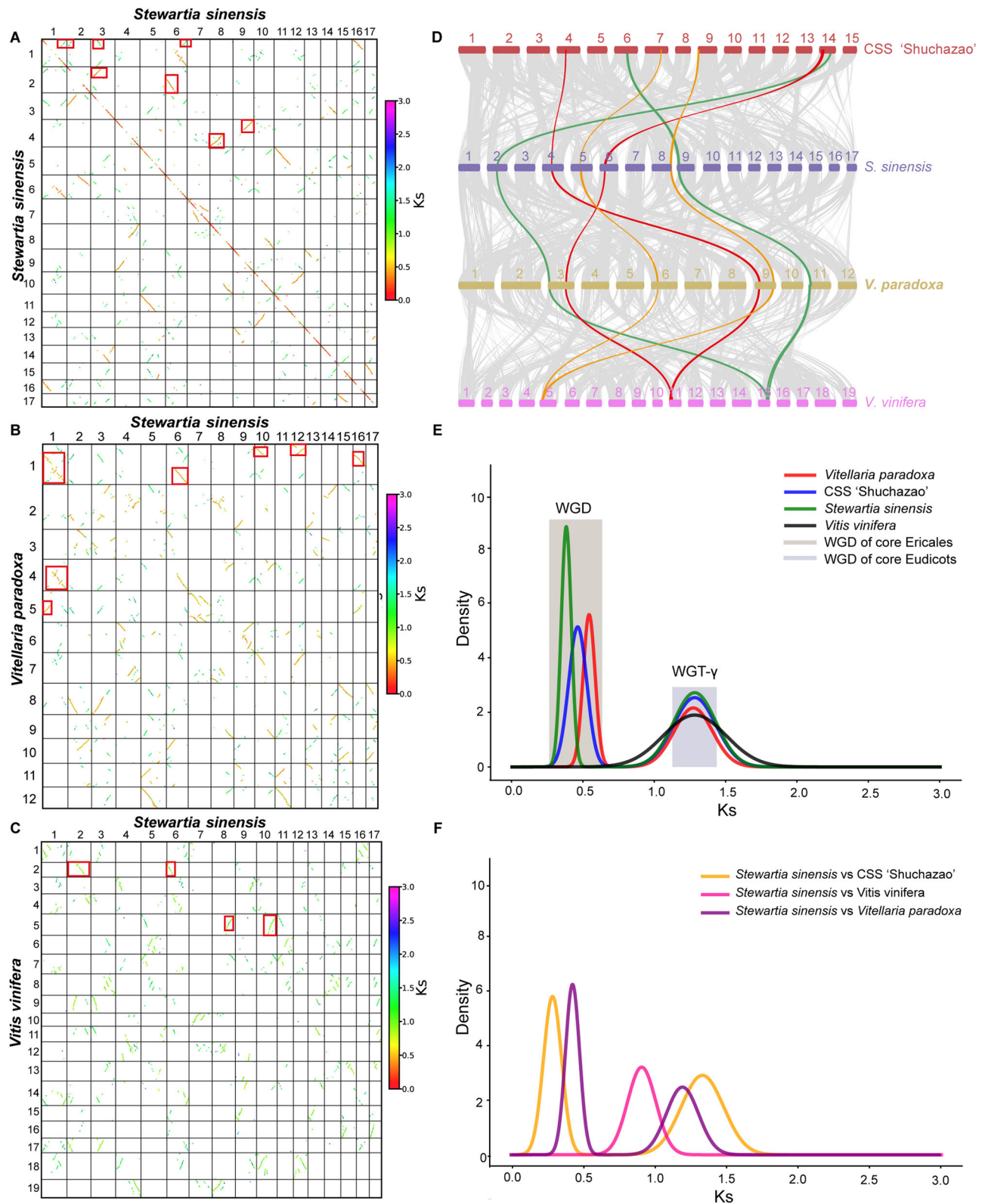
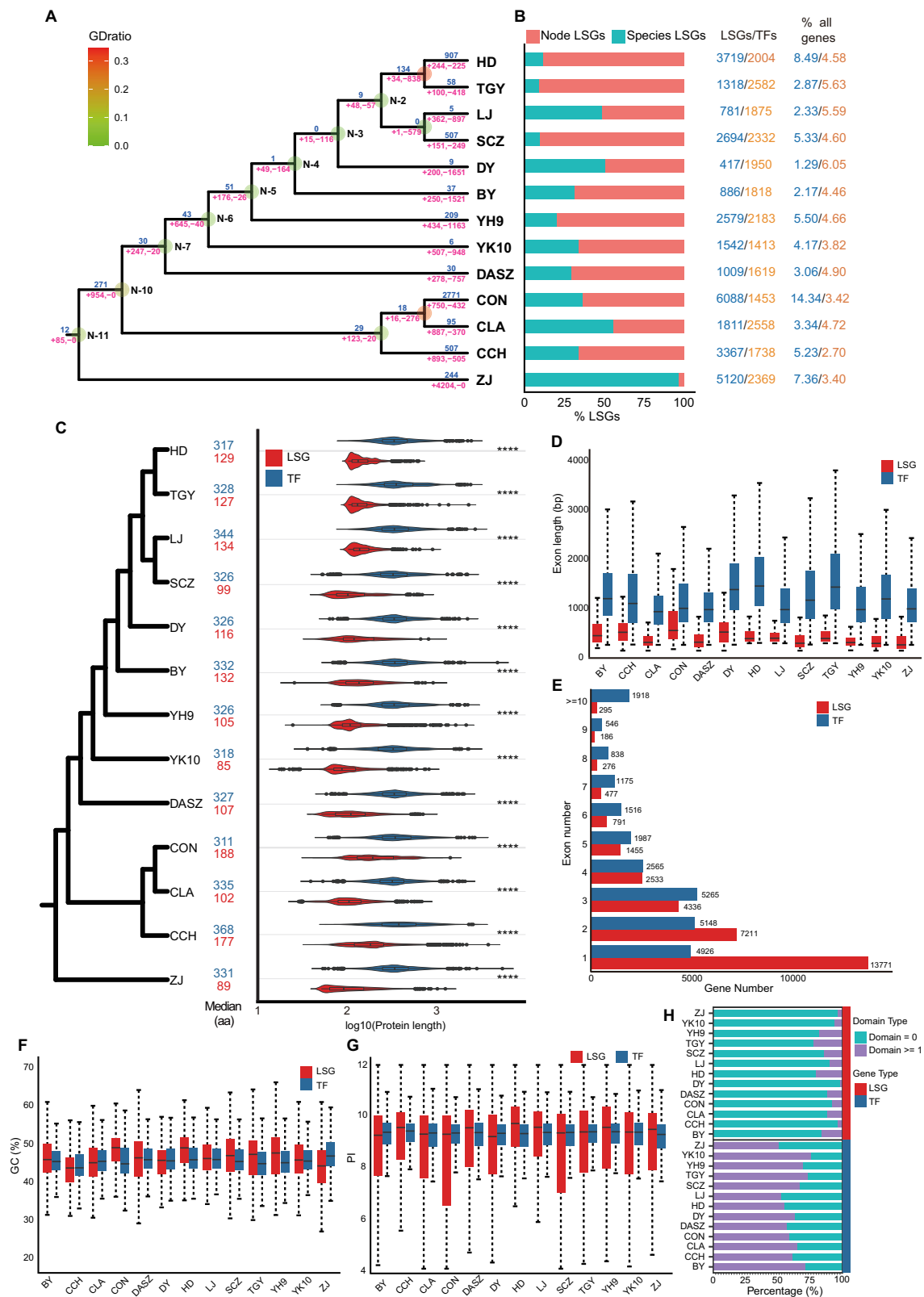


Fig. 4 | Identification of genome-wide duplication events in Theaceae. **A** Dot-plot of syntenic blocks in *S. sinensis*: This panel displays a dot-plot representation of syntenic blocks within the *S. sinensis* genome. **B** Synteny between *S. sinensis* and *V. paradoxa*. **C** Synteny between *S. sinensis* and *V. vinifera*. **D** Conserved gene orders among *S. sinensis*, CSS 'Shuchazao', *V. paradoxa*, and *V. vinifera*. **E** K_s Distribution

in paralogous gene pairs across genomes of *S. sinensis*, CSS 'Shuchazao', *V. paradoxa*, and *V. vinifera*, post-correction. **F** K_s distribution in orthologous gene pairs in comparing the genomes of *S. sinensis* with CSS 'Shuchazao', *V. paradoxa*, and *V. vinifera*, respectively.



depicted in Supplementary Fig 9. A species-wise breakdown revealed diverse orphan gene counts: CSS 'Shuchazao' contained 2694 orphan genes, CSS 'Longjing' 781, CSS 'Tieguanyin' 1318, CSS 'Huagandan' 3719, CSS 'Duyunmaojian' 417, CSS 'Biyun' 886, CSS 'Yinghong' 2579, CSA 'Yunkang10' 1542, *C. DASZ* 1009, *C. lanceoleosa* 1811, *C. oleifera* var. *Nanyongensis* 6088, *C. chekiangoleosa* 3367, and *S. sinensis* 5120, as

enumerated in Supplementary Table 15. Notably, over 40% of species-specific orphan genes were discovered in *C. chekiangoleosa* and *S. sinensis*. In *C. oleifera* var. *Nanyongensis*, a significant number of orphan genes, 2771 gene families were observed to be duplicated (Fig. 5A). In contrast, the study identified a total of 25,884 transcription factor (TF) genes across the 13 Theaceae species, as documented in Supplementary Table 16

Fig. 5 | Distribution of orphan genes and transcription factors in Theaceae.

A Gene duplication and family dynamics: This panel shows gene duplication (GD) counts above the horizontal line, with “+” symbols indicating gained gene families, “-” symbols for lost gene families, and “N” marking nodes in the evolutionary species tree. **B** Proportion of orphan genes in ancestral nodes and species: The bar graph illustrates the proportion of orphan genes in ancestral nodes and individual species. The numbers on the right detail the count and percentage of orphan genes and TFs in each sample relative to the total gene count. **C** Protein sequence length distribution of orphan genes and TFs in 13 Theaceae plants. The left side presents phylogenetic relationships, the middle shows median protein sequence lengths for each group, and the right side depicts the distribution of protein length within each Theaceae plant. An independent sample t-test was performed, with **** $P < 0.0001$ indicating statistical significance. **D** Exon length distribution. **E** Exon numbers in orphan genes and TFs. The horizontal axis representing the number of genes and the vertical axis

showing the number of exons per gene. **F** GC content distribution. **G** Isoelectric point (PI) distribution. **H** Domain number distribution. The purple bar with the vertical red bar on the right indicates orphan genes with at least one domain. The purple bar with the vertical blue bar on the right indicates TFs with at least one domain. The light blue bar with the vertical red bar on the right indicates orphan genes without any domain. The light blue bar with the vertical blue bar on the right indicates TFs without any domain. SCZ: *Camellia sinensis* var. *sinensis* ‘Shuchazao’, LJ: *Camellia sinensis* var. *sinensis* ‘Longjing’, TGY: *Camellia sinensis* var. *sinensis* ‘Tieganyin’, HD: *Camellia sinensis* var. *sinensis* ‘Huangdan’, DY: *Camellia sinensis* var. *sinensis* ‘Duyun’, BY: *Camellia sinensis* var. *sinensis* ‘Biyun’, YH9: *Camellia sinensis* var. *assamica* ‘Yinghong9’, YK10: *Camellia sinensis* var. *assamica* ‘Yunkang10’, DASZ: *Camellia* DASZ, CLA: *Camellia lanceoleosa*, CON: *Camellia oleifera* var. *Nanyongensis*, CCH: *Camellia chekiangoleosa*, ZJ: *Stewartia sinensis*.

and Supplementary Data 3. The distribution included 2322 TFs in CSS ‘Shuchazao’, 1875 in CSS ‘Longjing’, among others, with a comprehensive breakdown provided in Supplementary Table 16. The average TF gene count per species was approximately 2000, showing a relatively stable presence, accounting for about 2–6% of the total gene numbers in these plants, as shown in Fig. 5B, Supplementary Table 15, 16, Supplementary Data 3.

Sequence characteristics of orphan genes and TFs of Theaceae

In the 13 Theaceae plants, the orphan genes exhibited shorter amino acid sequence lengths and exon lengths compared to the TFs, as shown in Fig. 5C, D and detailed in Supplementary Data 4, 5. Contrasting with TFs, over 65% of orphan genes in Theaceae had one or two exons. There were only 295 orphan genes with more than ten exons, while 1918 TFs contained more than ten exons, as depicted in Fig. 5E and Supplementary Data 4, 5. Additionally, a minimal proportion of orphan genes encoded detectable functional protein domains, in stark contrast to over 90% of TFs, which possessed at least one protein domain (Fig. 5F and Supplementary Table 17). Whereas orphan genes in cultivated tea plants had higher GC content ratios than TFs (Fig. 5G, Supplementary Data 4, 5), orphan genes in *S. sinensis* displayed lower GC content ratios than TFs. The isoelectric points of most orphan genes in Theaceae plants are higher than those of TFs, as indicated in Fig. 5H and Supplementary Data 4, 5.

Evolutionary characteristics of orphan genes in Theaceae

Across the evolutionary timeline of Theaceae plants, including *C. DASZ*, *CSA* ‘Yunkang10’, and *CSS* ‘Shuchazao’, the amino acid length of orphan genes progressively decreases, which is consistent with the finding that orphan genes in these representative species possess fewer exons compared to TFs (Fig. 6A, Supplementary Data 4, 5). An analysis of GC content in orphan genes and TFs at different evolutionary stages showed an initial decrease followed by a gradual increase in GC content, eventually approximating the levels in TFs across these species (Fig. 6A). At the node N-11 of Theaceae, we identified 85 orphan gene families and 1149 TF gene families with TFs being about ten times more abundant. Across the 13 Theaceae plants, the number of orphan gene and TF gene families with at least four gene members was identified as 1240 and 1368, respectively (Fig. 6B). Our analysis revealed that at nodes N-10, N-9, N-7, N-6, and N-5, orphan genes gained 954, 123, 247, 645, and 176 gene families, respectively, while TFs acquired fewer gene families at the same nodes (Fig. 6B). Conversely, at nodes N-8, N-4, N-3, N-1, and N-0, orphan genes experienced a notable increase in lost gene families, surpassing the number lost by TFs by four times. Furthermore, orphan genes within the 13 Theaceae plants exhibited a substantially higher count of both gain and lost gene families compared to TFs.

Analysis of the nonsynonymous to synonymous mutation ratio (Ka/Ks) of the orthologous orphan gene pairs and TFs in *S. sinensis* were compared with eight representative species in cultivated tea groups at N11 (Fig. 6C, Supplementary Fig. 10A, D and Supplementary Data 6, 7). This method was similarly applied to the N10 and N6 (Fig. 6C, Supplementary Fig. 10B, C, E, F). Results suggested that orphan genes have undergone stronger positive selection, indicating higher evolutionary changes and potentially novel functional

alterations. To identify the expression pattern of each phylostratum, we used transcriptome data from eight tissues of CSS ‘Shuchazao’, including apical bud, flower, fruit, young leaf, mature leaf, old leaf, root, and stem (Fig. 6D, Supplementary Data 8). A high transcriptome age index (TAI) value indicates an evolutionarily young transcriptome, while a low TAI value signifies an evolutionarily ancient transcriptome. Phylotranscriptomic analysis combining gene age and expression information revealed generally higher TAI values for the flower than the other tissues and compared root tissue, more newly originated genes tend to function in flower, apical bud and fruits (Fig. 6D). We also identified 97, 1225, 277, 152, and 2428 orphan genes that have undergone GD events at their respective nodes (Fig. 6E and Supplementary Table 18), which highlights the important role of GD in the production of orphan gene in Theaceae and its potential influence on species diversification. Further examination of the gene duplication types for orphan genes showed that syntenic and tandem duplications contributed to less than 10% of the total number of orphan genes generated from gene duplications (Fig. 6E and Supplementary Table 18).

Expression characteristics of orphan genes and TFs in Theaceae

Gene expression data from CSS ‘Shuchazao’, *CSA* ‘Yunkang10’, *C. DASZ*, and *C. chekiangoleosa* were gathered across 21 different tissues and treatments (<http://tpia.teaplants.cn/>). For CSS ‘Shuchazao’, the expression levels of orphan genes tended to be lower than those of TFs (Fig. 7A and Supplementary Data 9, 10). In CSS ‘Shuchazao’, *CSA* ‘Yunkang10’, *C. DASZ*, and *C. chekiangoleosa*, 995, 680, 544 and 1183 orphan genes were detected expression, representing 36.93%, 44.10%, 53.91%, and 35.14% of the total orphan gene count, respectively (Fig. 7B–E and Supplementary Table 19). For TFs, 2295, 1400, 1606, and 1644 genes were expressed, constituting 98.84%, 99.08%, 99.20%, and 94.59% of the total TF count (Fig. 7B–E and Supplementary Data 11, 12). This demonstrates that over 60% of the orphan genes in the four Theaceae species are expressed in at least one tissue or treatment (Supplementary Data 11, 12).

To evaluate the expression breadth, we used a tissue specificity index value (τ), where higher τ values indicate stronger tissue-specific expression and lower values suggest broader expression. Orphan genes in these species typically had higher τ values compared to TFs (Fig. 7F and Supplementary Data 11, 12). Subcellular localization prediction for orphan genes and TFs in Theaceae indicated that over 80% of the TFs were predicted to be nuclear, with a minority located in other compartments (Fig. 7G and Supplementary Data 13). In contrast, less than 35% of orphan genes were predicted to be nuclear, with the remainder anticipated to localize in various cellular areas, including the cytoplasm, chloroplasts, mitochondria, and extracellular matrix. Compared to the GO analysis of TD genes (Supplementary Fig. 7), the orphan genes were mainly enriched in broader pathways such as regulation of primary metabolic processes, structural molecule activity, and binding, without specific enrichment in metabolite biosynthesis or degradation pathways (Fig. 7H–J, Supplementary Data 14–16). For example, *GWHTABKB010367*, *GWHTABKB011882*, *GWHTABKB012379*, *GWHTABKB012514*, *GWHTABKB013370*, and *GWHTABKB030837* were functionally enriched in terms such as zinc binding and structural molecule

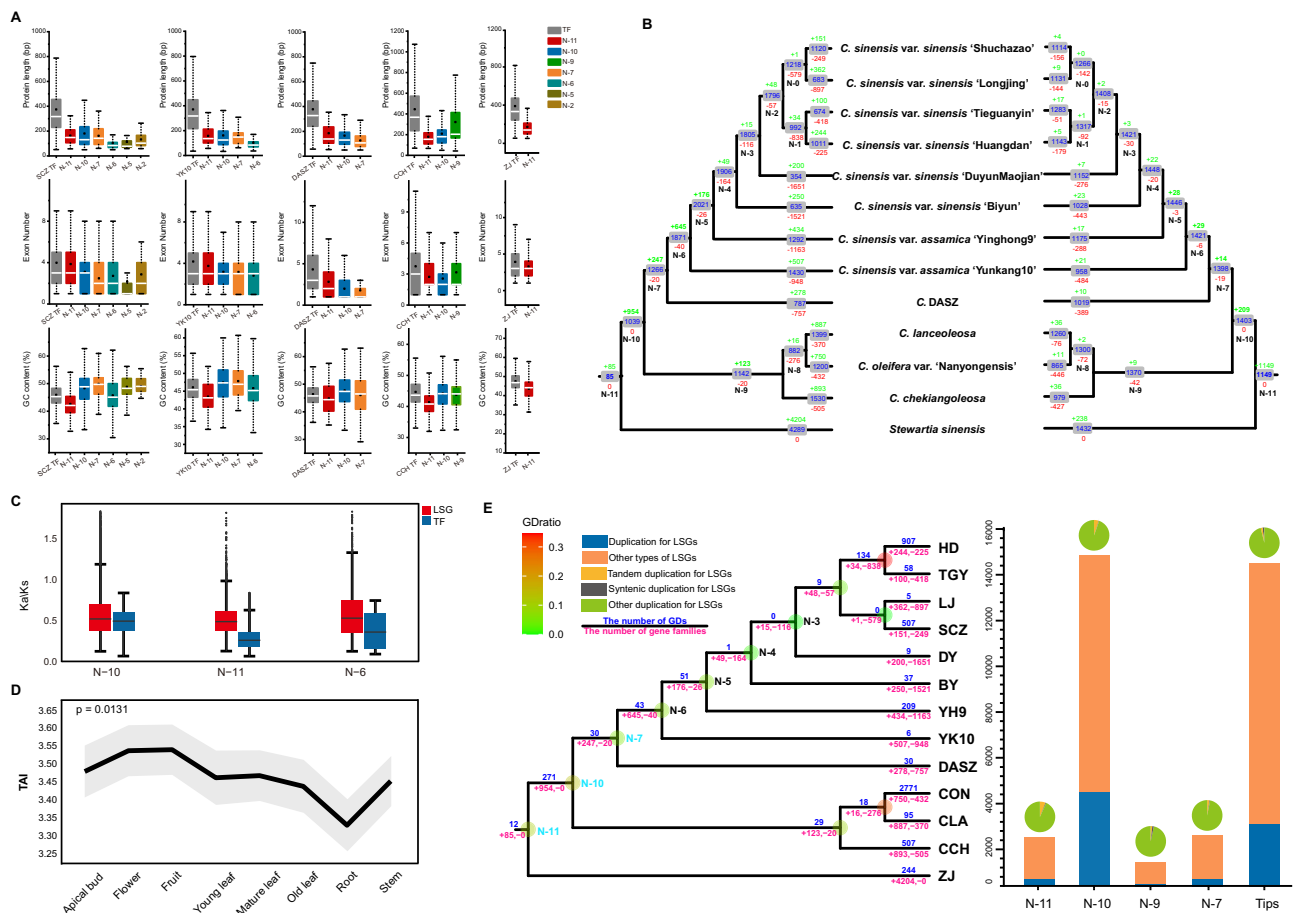


Fig. 6 | Evolutionary characteristics of orphan genes in Theaceae. **A** Protein sequence, exon number, and GC content distributions for orphan genes and TFs in CSS 'Shuchazao', CSA 'Yunkang10', C. DASZ, C. chekiangoleosa, and S. sinensis at different evolutionary nodes. Gray bars represent TFs identified in various Theaceae plants, while other colors indicate orphan genes at different nodes. **B** Gene family dynamics of orphan genes and TFs in 13 Theaceae plants. The left side represents gene family dynamics in orphan genes, while the right side shows TFs. The numbers in the box indicate the count of gene families, with "+" for gains, "-" for losses, and "N" for nodes in the evolutionary tree. **C** Ka/Ks values comparison in orthologous gene pairs between orphan genes and TFs at different nodes, namely N-11 (ancestor node of Theaceae), N-10 (ancestor node of the *Camellia* genus), and N-6 (ancestor node of cultivated tea). **D** TAI (Transcriptome Age Index) profiles. A high TAI value indicates that the sample expresses a high number of evolutionarily young genes. **E** Orphan gene duplication types: The left side shows the phylogenetic relationships of 13 Theaceae plants. Above the horizontal line, the number of gene duplications (GDs) in orphan genes at ancestral nodes is indicated. The "+" symbol

denotes gene families that have been gained, while the "-" symbol represents those that have been lost in orphan genes. The bar chart on the right side displays the number of orphan gene duplications and other types of orphan genes at different nodes. Blue represents the number of orphan gene duplications, while orange indicates other types of orphan genes. The pie chart demonstrates the proportion of orphan genes in different duplication types: yellow for tandem duplication, gray for syntenic duplication, and green for other duplication types. "N" denotes nodes, and "Tips" indicates gene duplications occurring during the formation of the 13 Theaceae plants. Species abbreviations: SCZ: *Camellia sinensis* var. *sinensis* 'Shuchazao', LJ: *Camellia sinensis* var. *sinensis* 'Longjing', TGY: *Camellia sinensis* var. *sinensis* 'Tieguanyin', HD: *Camellia sinensis* var. *sinensis* 'Huangdan', DY: *Camellia sinensis* var. *sinensis* 'Duyun', BY: *Camellia sinensis* var. *sinensis* 'Biyun', YH9: *Camellia sinensis* var. *assamica* 'Yinghong9', YK10: *Camellia sinensis* var. *assamica* 'Yunkang10', DASZ: *Camellia* DASZ, CLA: *Camellia lanceoleosa*, CON: *Camellia oleifera* var. *Nanyongensis*, CCH: *Camellia chekiangoleosa*, ZJ: *Stewartia sinensis*.

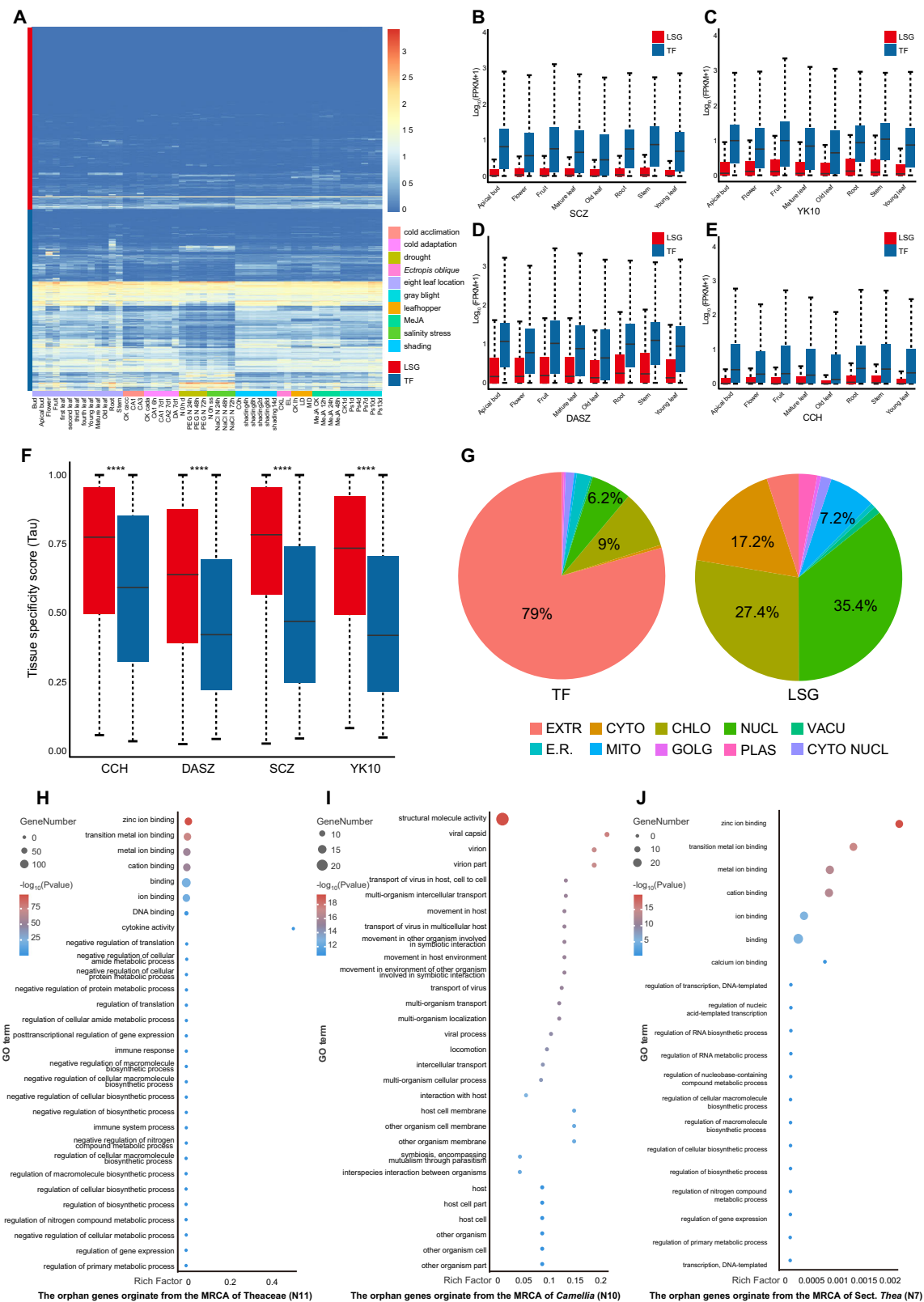
activity. Furthermore, these orphan genes were found to be expressed in young leaves, fruits, and stems of tea plants (Supplementary Data 9). These findings provide insights into the potential functional significance and tissue-specific expression of a set of previously uncharacterized genes in tea plants. The functional enrichment and tissue expression data point to these orphan genes having important biological roles worthy of further investigation.

Discussion

High-quality genome of *S. sinensis* contributes to the identification of orphan genes within Theaceae

The rapid advancement of sequencing technology has facilitated the decoding of numerous plant genomes. Particularly notable is the sequencing of genomes from basal angiosperm species, including water lily and *Amborella trichopoda*^{34,35}, has provided valuable insights into the early evolution of angiosperms and subsequent plant evolution.

Such genomic investigations have enhanced our understanding of the classification and functional roles of pivotal genes within angiosperms. On the other hand, the acquisition of high-quality genomes from basal species is critically important for exploring the evolutionary dynamics and facilitating the identification of orphan genes in plants^{73,74}. The publication of the *Cercis chinensis* genome, a foundational species in the legume family, has provided a basis for investigating the origin and evolutionary models of orphan genes in legumes⁷⁴. Analyses revealed that the de novo originated new genes differ significantly from established genes in coding sequence length, exon number, GC content, and tissue expression patterns, consistent with the new genes identified in bamboo³⁶. The new genes identified in bamboo preferentially expressed in rapidly developing shoots, indicating their potential role in morphogenesis during fast growth³⁶. To date, 12 genomes in Theaceae have been published, primarily focusing on the tea or Oil-Camellia plants^{48–51,53,54,75–78}. The *Stewartia* represents the



earliest divergence clade within the Theaceae³. We successfully assembled a high-quality genome for *S. sinensis*, a crucial reference genome for understanding gene family dynamics and the evolutionary origins of orphan genes in ancestors of Theaceae and their sub-lineages. We identified 31,331 orphan genes, the characteristics

of these new genes, such as gene length and number of exons, are consistent with the new genes identified in *Cercis chinensis* and bamboo, with tandem duplication significantly contributing to the adaptive evolution and diversity of Theaceae, serving as a major mechanism driving the origination of orphan genes.

Fig. 7 | Expression characteristics and functional prediction of orphan genes in Theaceae. **A** Gene expression in CSS ‘Shuchazao’ across various tissues and different environmental conditions. The horizontal axis lists different tissues and stress conditions including bud, flower, leaves of various ages, root, stem, and treatments like cold, drought, shading, pest stress, and infection. The vertical axis represents the expression levels of orphan genes and TFs. **B–E** Expression levels of orphan genes and TFs in eight different tissues for CSS ‘Shuchazao’ **B**, CSA ‘Yunkang10’ **C**, C. DASZ **D**, and C. *cheiangoleosa* **E**. **F** The tissue expression specificity index (Tau index) for CSS ‘Shuchazao’, CSA ‘Yunkang10’, C. DASZ, and C. *cheiangoleosa*. The Tau index quantifies gene expression specificity across tissues, with statistical significance indicated by an independent sample T-test (**** $P < 0.0001$). **G** Subcellular localization prediction of orphan genes and TFs in Theaceae plants.

The left and right sides represent TFs and orphan genes across 13 Theaceae plants. Various cellular components are indicated, including extracellular (EXTR), chloroplast (CHLO), cytoplasm (CYTO), cytoplasm_nucleus (CYTO NUCL), extracellular (E.R.), Golgi body (GOLG), mitochondria (MITO), nucleus (NUCL), plastids (PLAS), and vacuole (VACU). **H–J** GO enrichment analysis of orphan genes originated from the nodes N11 (the MRCA of Theaceae), N10 (the MRCA of *Camellia*), and N7 (the MRCA of Sect. *Thea*). The horizontal axis (rich factor) represents the ratio of differential genes under each metabolic pathway to all genes annotated in that pathway; the vertical axis indicates the enriched metabolic pathway. The size of each circle denotes the number of genes annotated to the corresponding GO item, and the color represents the significant level of enrichment results.

The most recent WGD- β in Theaceae was dated at approximately 91.06 Mya

The limitations of using transcriptomic data for identifying WGD events include the incomplete coverage of the genome⁷⁹, which may result in an incomplete representation of genomic duplications. This limitation restricts the comprehensive identification of WGD events, particularly for smaller-scale duplications that may be challenging to detect. Although previous studies have commonly recognized that all Theaceae members underwent the core eudicots whole-genome triplication (WGT- γ) event, the occurrence of subsequent WGD events in the MRCA of Theaceae, as well as the existence of specific WGD events within Theaceae, remain topics of ongoing debate due to the lack of genomic information from the early-diverging clade of Theaceae. Presently, there are three main viewpoints in this debate for WGD events related to Theaceae. One perspective suggests that following the WGT- γ event, Theaceae underwent a shared WGD- β event with kiwifruit and rhododendron^{46,49,54,80}. The position of this WGD event is still highly controversial. The second viewpoint, emerging from a genomic study of CSS ‘Shuchazao’, suggests an additional tea plant-specific WGD event that occurred ~30 to 40 Mya⁵³, termed as Cm- α ⁸¹, alongside the WGT- γ and WGD- β events^{53,81}. The third viewpoint posits that tea plants underwent a single tea plant-specific WGD event subsequent to the WGT- γ event, supporting the conclusion reached by Yang et al.^{75,82}. The challenge in resolving this debate is partly due to the limited availability of publicly accessible genomes within Theaceae, predominantly from the *Camellia* genus. The utilization of the tea plant as the sole representative species of Theaceae in most of the aforementioned studies led to inaccuracies in identifying WGDs at ancestral nodes of this family. Our study, incorporating the *S. sinensis* genome alongside CSS ‘Shuchazao’, *Vitis vinifera* and *Vitellaria paradoxa* genomes, aimed to investigate WGD events within the family. The *V. vinifera* genome has served as a valuable reference genome for evolutionary studies, primarily due to its ancestral eudicot chromosome structure and the absence of additional WGD events, except for the WGT- γ ⁸³. In previous studies investigating the WGD events in Theaceae, the genome of kiwifruit, the closest genus to *Camellia* in phylogenetic trees, was commonly used for Ks analysis. However, compared to *V. paradoxa*, the kiwifruit experienced two subsequent tetraploidization events following the WGT- γ . These events include the *Actinidia* recent tetraploidization occurring approximately 18–20 Mya and the ancient tetraploidization dating back to around 50–57 Mya^{84,85}. In addition, kiwifruit has undergone more substantial artificial selection as a significant fruit crop, which also resulting in a larger Ks peak value and faster evolutionary rate. This poses challenges for accurately identifying polyploidy events. In our study, we selected *Vitellaria paradoxa* from the Sapotaceae family as the reference genome to explore evolutionary patterns, given its closer relationship to the ancestral node of core Ericales (Fig. 3). We confirmed that Theaceae experienced only one round of WGD- β around 91.06 Mya since the WGT- γ , which is shared by other families in the order Ericales including Theaceae, Ericaceae, Clethraceae, Actinidiaceae, Roridulaceae, Primulaceae, Ebenaceae, Sapotaceae, Polemoniaceae, et al. (Figs. 3 and 4, Supplementary Fig. 3). WGD has been widely acknowledged as a significant driving force in the evolution of speciation, adaptation, and diversification⁸⁶. Emerging research suggests that polyploid plants with duplicated genomes exhibit enhanced

adaptability and improved tolerance to diverse environmental conditions, which may have contributed to their higher survival rate during the Cretaceous-Tertiary extinction event about 66 Mya^{87,88}. Our study indicates that despite the recent WGD event in Theaceae not aligning with the K-Pg boundary, the doubling of the genome played a crucial role in facilitating the differentiation and survival of Theaceae during the mass extinction event.

Tandem duplications contributing greatly to the adaptation and diversification of Theaceae

Duplicated genes offer greater opportunities for natural selection to shape novel functions⁸⁹. Tandem gene duplication, which involves two or more adjacent homologous genes, is a key source of single-gene duplication in plant genomes and plays a significant role in plant evolution and adaptation to environmental changes⁹⁰. In *Cymodocea nodosa* and *Potamogeton acutifolius*, additional recent lineage specific WGDs and tandem duplications contributed to further expansion of the hypoxia responsive genes as a potential adaptation to submergence⁹¹. In *Pinellia pedatisecta* and *Tamarix chinensis*, TDs contributed to the expansion of genes associated with stress responses^{92,93}. Based on our previous studies³ and the genomic information of the Theaceae species *S. sinensis* generated in this study, given the lack of Theaceae-specific WGD, we focused on the contributions of gene duplication and orphan gene origination to the diversification and trait formation within Theaceae. We found that at these three key evolutionary nodes, duplicated genes account for 10–30% of the total orphan genes (Fig. 6E). Notably, tandem duplications and segmental duplications contributed by synteny account for less than 10% of the duplicated genes.

It is interesting that functional enrichment analyses of the tandem duplication (TD) genes and orphan genes at the N11, N10, and N7 nodes, as well as in *S. sinensis* and CSS ‘Shuchazao’ (Supplementary Fig. 7, Fig. 7H–J), revealed that the TD genes at the N11 node were enriched in functions related to catalytic activity, mannitol dehydrogenase activity, chlorophyll catabolic process, recognition of pollen, and carbohydrate transporter activity, et al. (Supplementary Fig. 7). These functions, such as the key role of mannitol dehydrogenase in plant stress responses and the importance of chlorophyll catabolism and pollen recognition in plant life cycle, likely helped the Theaceae ancestors survive and reproduce in harsh environments. At the MRCA of genus *Camellia* node N10, the TD genes were more enriched in terpene synthase activity, flavonoid biosynthesis, and triglyceride lipase activity (Supplementary Fig. 7), which may have laid the foundation for the diversification and fast radiation of the economically important *Camellia* species, including the tea plant³. In contrast, the orphan genes were predominantly enriched in broader pathways such as regulation of primary metabolic processes, structural molecule activity, and binding, without specific enrichment in metabolite biosynthesis or degradation pathways (Fig. 7H–J). This difference from the previous findings by Xia et al. (2017), which reported lineage-specific expansion of genes related to flavonoid metabolism and enhancement of catechin production, terpene activation and stress tolerance in tea plants⁵⁴, may be due to the inclusion of the *S. sinensis* in our analysis and the use of more stringent orphan genes identification criteria, which better separated the TD genes from the orphan genes.

Materials and Methods

Plant material preparation and sequencing for *S. sinensis*

Sample collection and DNA extraction: fresh leaves were collected from a *S. sinensis* plant at Hangzhou Botanical Garden (Hangzhou City, Zhejiang Province; N30°16', E120°12', elevation 20–60 meters). Previous morphological studies that considered *S. gemmata* to be a synonym of *S. sinensis*⁵. Genomic DNA was extracted from these leaves using an enhanced CTAB method⁹⁴.

Next-generation sequencing library construction and sequencing: DNA samples underwent random fragmentation using the Covaris ultrasonic crusher. The Illumina sequencing library was prepared using the Nextera DNA Flex Library Prep Kit (Illumina, San Diego, CA, USA) following a series of steps including the DNA fragments with a target insertion fragment size of 150 bp were subjected to terminal repair, A-tail addition, sequencing splice addition, purification, and PCR amplification. After qualified library detection, high-throughput sequencing was performed on the DNBSEQ-T7 platform.

Third-generation sequencing library construction and quality assessment: large fragments of DNA (greater than 15 Kb) were enriched and purified using magnetic beads. This process included the repair of damaged ends and polishing of the fragmented DNA. After purification, A-tailing was performed on both ends of the DNA fragments. Sequencing adapters from the SQK-LSK109 kit were then ligated to prepare the library. The constructed DNA library's concentration was precisely measured using Qubit. Sequencing was conducted on the PromethION P48 sequencer (Oxford Nanopore Technologies, Oxford, UK) after loading the purified library onto R9.4 Spot-On Flow Cells.

Hi-C data sequencing: the sample was subjected to formaldehyde-induced cross-linking, followed by digestion with the restriction enzyme *DpnII* (New England Biolabs, MA, USA). This step generated cohesive ends adjacent to the cross-linking sites. Subsequent steps included non-homologous end joining, circularization, DNA purification, capture, and library quality assessment. The final sequencing was performed on an Illumina NovoSeq 6000 platform (Benagen Technologies, Hubei, China), generating paired end reads of 150 base pairs each.

Hi-C assisted chromosomal-level genome assembly and scaffolding

The raw data of Hi-C sequencing were processed using HICUP (v0.8.0) to extract valid chromatin interaction pairs⁹⁵. In the alignment process, reads that did not uniquely align to the reference genome were discarded. Additionally, invalid read pairs and duplicates resulting from PCR amplification were filtered out. This filtering step ensured only valid interaction pairs were retained for subsequent analysis. The initially assembled contigs were scaffolded using 3D-DNA and Juicer software^{96,97}, leveraging these valid interaction pairs to refine the draft genome sequence. This methodology enabled the effective anchoring of contigs to chromosomes, culminating in the assembly of the *S. sinensis* genome at the chromosomal level.

Genome annotation of *S. sinensis*

Employing a de novo approach, the software RepeatModeler was used to predict model sequences based on the genome sequence from the *S. sinensis* genome⁹⁸. Additionally, LTR_FINDER software was utilized to predict LTR (Long Terminal Repeat) sequences⁹⁹. The LTR_retriever software was utilized to conduct a de-redundancy process on sequences predicted by LTR_FINDER¹⁰⁰. The RepeatMasker subroutine RepeatProteinMask was used for predicting transposable element (TE) protein type repeat sequences. Gene structure prediction was performed using a comprehensive approach, combining transcriptome, homology, and de novo predictions.

For the next and third generation sequencing data, the genome was compared using the software HISAT2 v2.1.0 and minimap2 v2.17^{101,102}. The aligned data in BAM format were further processed using Stringtie v2.1.4 with the parameter “-a p15” to reconstruct transcripts¹⁰³. TransDecoder v5.1.0 software was utilized to predict coding frames within sequence. Homologous predictions involved five related species including *CSS* ‘Tieganyin’, *C. DASZ*, *C. lanceoleosa*, *A. chinensis* and *V. darrowii*, with protein sequences compared against the genome using the tBLASTN algorithm.

Results were further utilized for transcript and coding predictions with Exonerate v2.4.0 (<https://github.com/nathanweeks/exonerate>). Gene predictions were performed using Augustus v3.3, Genscan v1.0, and GlimmerHMM v3.0.4, with Genscan (<http://genome.ucsc.edu/cgi-bin/hgTrackUi?g=genscan>) specifically used for the prediction process^{104–106}. The MAKER v2.31.10 software integrated and consolidated gene sets from these methods.

Sequence and motif similarities were used to annotate the gene function of *S. sinensis*. Functional information and metabolic pathway associations of protein sequences were analyzed using diamond BLASTp v2.0.11.149¹⁰⁷ against databases such as UniProt, NR, and KEGG¹⁰⁸. Protein domain architecture and transmembrane regions were predicted using InterProScan v5.52-86.0, querying protein sequences against secondary databases within InterPro, including CDD, Gene3D, Hamap, Panther, Pfam, Phobius, Pirsf, Pirsr, Prints, Prosite, Sfld, Smart, Superfamily and HMMER v3.3.2 were used for domain prediction to identify conserved sequences, motifs and domains of proteins¹⁰⁹. tRNAscan-SE v1.23 (parameter -q) was utilized to search for tRNA sequences based on their structural characteristics¹¹⁰. The ncRNA sequences were annotated with INFERNAL v1.1.2 with the parameters “-cut_ga -rfam -nohommonly -cpu 15” using Rfam database¹¹¹.

Integration of genomic and transcriptomic datasets: assembly, annotation and BUSCO completeness assessment

In this study, we compiled a comprehensive collection of 181 publicly available angiosperm datasets, comprising 31 genomic and 150 Theaceae transcriptomic datasets. This collection, augmented by our high-quality genome assembly of *S. sinensis*, is crucial for the accurate identification and characterization of orphan genes. The genomic data included 13 high-quality datasets from Theaceae and 18 genomes from outgroups taxa such as Ericaceae, Clethraceae, Actinidiaceae, Roridulaceae, Primulaceae, Ebenaceae, Sapotaceae, and Polemoniaceae. Additionally, we incorporated genomes from Nyssaceae and Hydrangeaceae (Cornales)^{112–115}, Brassicaceae (Brassicales)¹¹⁶, Vitaceae (Vitales)⁸³, and *Aquilegia coerulea*¹¹⁷ as a representative of basal eudicots. The genome of *A. trichopoda*, representing basal angiosperms, was also included³⁴. Transcriptomic data mainly comprised 133 samples from the tribe Theaeae, nine samples from the Tr. Gordoneae, and eight samples from the Tr. Stewartiae. Transcriptomic data were from the National Center for Biotechnology Information (NCBI, <https://www.ncbi.nlm.nih.gov>).

The software Trimmomatic (v0.39)¹¹⁸ was used to perform filtering on a dataset consisting of 149 publicly available transcriptomes and shallow genome data samples (referred to as “Raw data”) (Supplementary Data 1). *Camellia sinensis* var. ‘Huangjinju’ is represented by a CDS and PEP file that was directly downloaded. The process involved removing low-quality reads using parameters as “LEADING:10 TRAILING:10 SLIDINGWINDOW:4:20 MINLEN:36”. De novo assembly of all transcriptomes was conducted with Trinity (v2.11.0)¹¹⁹, constructing contigs from the raw transcriptomic data. The software TransDecoder (v5.5.0) (<https://github.com/TransDecoder>, accessed in August 2023) was utilized for the prediction of coding sequence (CDS) regions. CD-HIT program (v4.8.1) was employed to reduce redundancy within each assembly, with a parameter of “-c 0.98”¹²⁰. The software BUSCO (v5.2.2) was used to assess the completeness of gene annotations for each sample using the eudicots_odb10 database, specifically designed for eudicot plant species⁷⁹. Details regarding the transcriptomes generated in this study (Supplementary Data 1), as well as the assembly completeness assessment using BUSCO, are provided in Supplementary Data 1 and 2.

Phylogenetic analysis

We used our customized script to convert the nucleic acid sequences into their corresponding amino acid sequences (<https://github.com/Linlin1983/Stewartia-sinensis> or <https://doi.org/10.5281/zenodo.14511253> as cds2prot.pl). For a comprehensive protein sequence comparison, the software DIAMOND was utilized to perform all-against-all BLASTP analysis on the amino acid sequences¹⁰⁷, with a default E-value 1e-5 for significance. The Markov Clustering Algorithm (MCL v14.137, <https://micans.org/mcl/>) was employed to cluster the gene pairs obtained from the BLASTP comparison¹²¹. This algorithm facilitated the grouping

the gene pairs into distinct clusters based on their similarity bit scores for each paired homolog.

Multiple sequence alignment of the clustered amino acid sequences was conducted using MAFFT v7.487¹²². The aligned amino acid sequences were then converted into nucleic acid sequences using PAL2NAL v13 software¹²³. To remove poorly aligned regions and enhance alignment quality, trimAl v1.2 software was employed with the parameters “automated1” to improve the overall alignment quality¹²⁴.

IQ-TREE v2.14-b was employed to construct a maximum likelihood (ML) gene family tree¹²⁵. The best model for tree construction was automatically determined by the software ModelFinder¹²⁶ for accurate phylogenetic inference, and 1000 bootstrap replicates were performed to validate the robustness of tree topology. Finally, ASTRAL-Pro¹²⁷ was employed to reconstruct the coalescent species tree.

Estimation of the divergence time and dating the Core-Ericales WGD

For divergence time estimation, we selected two fossil calibration points and two secondary calibration points as divergence time markers. The fossil calibration points chosen were 125–247.2 million years (Myr)^{35,128–131}, representing the crown group node of Angiosperms¹³² and the earliest tricolpate pollen fossils, marking the stem node of eudicots around 125 Myr¹³². The secondary calibration points were 79.8–102.5 Myr for the stem node of Theaceae and 39.6–74.7 Myr for the crown node of Theaceae⁴. The MCMCTree program within PAML¹³³ software was utilized for molecular clock analysis, using 1030 orthologous groups (OGs) with at least 90% coverage and gene lengths over 800 bp¹³³. This analysis incorporated the auto-correlation model, the General Time Reversible (GTR) site substitution model, and prior probabilities. Markov Chain Monte Carlo (MCMC) sampling estimated the posterior distribution of node ages, discarding the initial 200,000 generations and sampling every 20 generations over a total of 500,000 generations. Convergence was checked by repeated analysis and a sufficient sample size was examined.

The ages of WGD detected in this study were estimated based on the assumption of a constant rate of synonymous mutation accumulation. If a WGD event is flanked by two species divergence events in the species tree, the upper limit of the WGD age (denoted as T^{prior}) was set at the time of species divergence preceding the WGD, and the lower limit (denoted as T^{post}) at the divergence time following the WGD. The emergence times of lineages on the species tree, derived from a recent study using fossil data and computational estimates of chloroplast genes¹³⁴, determined T^{prior} and T^{post} for each WGD event. The time of a specific WGD event, T^{wgd} , was determined using the following function:

$$T^{\text{wgd}} = T^{\text{post}} + \frac{Ks^{\text{wgd}} - Ks^{\text{post}}}{Ks^{\text{prior}} - Ks^{\text{post}}} \times (T^{\text{prior}} - T^{\text{post}})$$

In the context of WGD events, Ks values play a critical role. Ks^{prior} and Ks^{post} refer to the highest synonymous substitution (Ks) values observed between orthologs of two diverging lineages before and after the WGD event, respectively. These values were derived by conducting all-against-all BLAST analysis to compare reciprocal best matched genes between paired species. Conversely, Ks^{wgd} represents the average Ks value across all gene duplicates resulting from the specific WGD event being studied.

Detection of gene duplication events in different Theaceae lineages

To predict gene duplication events, we conducted gene and species tree reconciliation using the Tree2GD tool¹³⁵. This approach allowed us to reconcile gene tree data with the species tree, facilitating the identification of gene duplication events within the evolutionary context. Additionally, we assessed collinearity both between and within species using the WGDI software with default parameters¹³⁶. This software analyzes genomic collinearity, which is crucial for understanding the conservation and diversification of gene order among different species or within the same species in the context of geological times.

Synonymous substitution rates (Ks) correction

We performed the Ks correction to accurately determine the origin of core Ericales WGD¹³⁷. This approach assumes a normal distribution for the evolutionary rate Ks values. Species A was selected as the reference to correct the Ks values of species B.

If $X_A \sim N(\mu_A, \sigma_A^2)$ represents the distribution of Ks values in species A, and the duplicated gene pairs in species B follow $X_B \sim N(\mu_B, \sigma_B^2)$, then the correction coefficient is given by:

$$\lambda_B = \frac{1}{1+r}; r = \frac{\mu_B - \mu_A}{\mu_A}$$

The corrected value of X_B , denoted as:

$$X_{B-\text{correction}} \sim N(\lambda_B \mu_B, \lambda_B^2 \sigma_B^2)$$

If there is another species C, there is:

$$X_{BC-\text{correction}} \sim N(\lambda_{BC} \mu_{BC}, \lambda_{BC}^2 \sigma_{BC}^2)$$

KEGG and GO analysis

The KEGG and GO enrichment analysis were conducted by the OmicShare tool platform, which can be accessed online at <https://www.omicshare.com/tools/home/report/koenrich.html>, and <https://www.omicshare.com/tools/home/report/goenrich.html>, respectively.

Identification of orphan genes and transcription factors (TFs) in Theaceae

In this study, the CDS sequences of 31 data sets were translated into amino acid sequences using the cds2prot.pl (uploaded to <https://github.com/Linlin1983/Stewartia-sinensis> and <https://doi.org/10.5281/zenodo.14511253>). These amino acid sequences were subjected to all-against-all comparison using Diamond software with an E-value threshold of $1e-5$ ¹⁰⁷. PhyloMCL was employed⁷² to cluster the genes by analyzing the based BLASTp bit scores, resulting in the identification of orthologous clusters or gene families across the 31 data sets. The dolloparsimony module within the Tree2GD software package was utilized to identify gene family gain and loss events¹³⁵. Gene families obtained from ancestral nodes, internal branches, and individual species were subsequently considered as candidate gene sets for specific genes in Theaceae. Candidate genes underwent sequential comparison against multiple databases including nonredundant protein sequences (NR), nonredundant nucleotide sequences (NT), SWISS-PROT (a manually annotated and reviewed protein sequence database), 1000 Plant transcriptomes initiative (OneKP) databases, and a collection of 150 Theaceae transcriptomic assembly gene sets. Candidates not matching any entries in these databases were designated as Theaceae orphan gene set one. This comparison process involved species classification, exclusion of non-Theaceae genes, and rejection of matches with Cyanobacteria. Additionally, the retrieved genes needed to meet sequence similarity criteria across all databases, with a threshold of at least 70% similarity, and each gene required a minimum of five matches with other genes (Supplementary Fig. 9). These criteria helped form gene set two. The cumulative total of these two gene sets represents the orphan genes identified within Theaceae.

Analysis of gene sequence characteristics of Theaceae

The orphan genes of 13 samples within Theaceae were analyzed, focusing on their proportion within the total gene count in each genome. Various characteristics such as amino acid sequence length, exon number and length, and GC content were determined. The isoelectric point (PI) of the encoded proteins was calculated using the software TBtools¹³⁸. Additionally, the Pfam database was employed to identify protein domains within the orphan genes and TFs of Theaceae and the number of protein domains was ascertained. Data visualization was performed using the software Origin for effective presentation of the findings¹³⁹.

To ascertain the reliability and precision in orphan gene identification, we employed a comprehensive approach incorporating public high-quality Theaceae genomes, including basal species *S. sinensis*, sequenced and assembled in this study. Additionally, we integrated genome data from 18 representative outgroup species, selected based on data quality, to enhance the robustness of our analysis. In the process of identifying orphan genes in Theaceae, a total of 192,751 genes from 56,150 gene families were identified, originating from ancestral nodes, internal branches, and the species themselves. These genes formed the candidate gene pool for orphan genes in Theaceae (Supplementary Fig. 9, and Fig. 5A).

Transcriptome age index calculation

The transcriptome age index (TAI) for each tissue and developmental stage was calculated using the myTAI package¹⁴⁰. The TAI value for a sample is defined as the weighted mean of the phylostratum rank (N_i) of gene i , weighted by its expression value (e_{is}) in the transcriptome of sample s , where n represents the total number of genes analyzed. The higher TAI value reflects that the genes from sampled transcriptome are broadly expressed in the evolutionarily younger nodes in the species tree¹⁴¹.

$$TAIs = \frac{\sum_{i=1}^n e_{is} \cdot N_i}{\sum_{i=1}^n e_{is}}$$

Analysis of gene expression characteristics of Theaceae

To access expression data of CSS 'Shuchazao', CSA 'Yunkang10', C. DASZ, and *C. chekiangoleosa* at different developmental stages and habitats, we used the Tea Plant Information Archive (TPIA) database (<http://tpia.teaplants.cn/>). For predicting the subcellular localization of TF protein sequences specific to Theaceae, the WoLF PSORT online website tool (<https://wolfpsort.hgc.jp/>) was employed. Additionally, to investigate the gain and loss of orphan genes and TFs within the Theaceae gene family evolution, the Tree2GD software was utilized¹³⁵.

The tissue expression specificity index, known as the Tau index (τ), was employed to quantify the extent of gene expression across various tissues¹⁴². The Tau index (τ) is calculated from normalized expression levels of genes (x_i) in different tissues relative to the maximum expression level observed across all tissues. The Tau index (τ) ranges 0 to 1, with $\tau = 0$ indicating broad expression across tissues and $\tau = 1$ highly indicating tissue-specific expression. The formula for the calculation of Tau index (τ) is as follows:

$$\tau = \frac{\sum_{i=1}^N (1 - \hat{x}_i)}{N - 1}; \hat{x}_i = \frac{x_i}{\max(x_i)}; 1 \leq i \leq N$$

Here, N represents the total number of tissues analyzed. This index provides a quantitative measure of gene expression specificity, aiding in the understanding of gene function and regulation in different tissue types.

Gene selection pressure analysis

To compare orphan gene with TF orthologous gene pairs at different phylogenetic nodes, multiple sequence alignments were conducted using the MAFFT software¹²². These alignments provided the basis for subsequent analyses of the evolutionary dynamics among the studied species. For computing the comparison results of the matched orthologous gene pairs, the KaKs_Calculator tool was utilized¹⁴³. This tool applied the Nei-Gojobori (NG) method to determine the ratio of K_a to K_s substitutions¹⁴³. This ratio, known as the K_a/K_s ratio, is a critical indicator of evolutionary selection pressure as: a K_a/K_s ratio greater than 1 suggests positive selection, less than 1 indicates purifying selection, and a ratio around 1 implies neutral evolution.

Statistics and reproducibility

Detailed information regarding the statistical analyses conducted is provided in the Methods section. All analyses can be replicated by accessing the associated data available in the Data Availability statement.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The comprehensive datasets for *S. sinensis*, encompassing Oxford Nanopore long-read and short-read whole-genome sequencing, Hi-C, and RNA-seq have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRR31837436, SRR31842164, SRR31850622 and SRR31853498), under the BioProject accession number PRJNA1060471. The resulting genome assembly sequences along with gene annotations have been deposited to CNSA under accession CNA0146085. Additional data that support the findings of this study are available from the corresponding author upon a reasonable request.

Code availability

Source code for analysis is available at Github <https://github.com/Linlin1983/Stewartia-sinensis> and <https://doi.org/10.5281/zenodo.14511253>.

Received: 12 February 2024; Accepted: 13 January 2025;

Published online: 03 March 2025

References

1. Glumac S. P., *The World Flora Online*, <https://www.worldfloraonline.org/> (2023).
2. Rose, J. P. et al. Phylogeny, historical biogeography, and diversification of angiosperm order Ericales suggest ancient Neotropical and East Asian connections. *Mol. Phylog. Evol.* **122**, 59–79 (2018).
3. Cheng, L. et al. Phylogenomics resolves the phylogeny of Theaceae by using low-copy and multi-copy nuclear gene makers and uncovers a fast radiation event contributing to tea plants diversity. *Biology* **11**, 1007 (2022).
4. Yu, X. Q. et al. Insights into the historical assembly of East Asian subtropical evergreen broadleaved forests revealed by the temporal history of the tea family. *New Phytol* **215**, 1235–1248 (2017).
5. Ming, T. L. & Bartholomew B. *Flora of China*. **12**, 424–429 (2007).
6. Spongberg, S. A. A review of deciduous-leaved species of *Stewartia* (Theaceae). *J. Arnold. Arbor.* **55**, 182–214 (1974).
7. Lin, H.-Y., Gu, K.-J., Li, W.-H. & Zhao, Y.-P. Integrating coalescent-based species delimitation with ecological niche modeling delimited two species within the *Stewartia sinensis* complex (Theaceae). *J. Syst. Evol.* **60**, 1037–1048 (2022).
8. Prince, L. M. Circumscription and biogeographic patterns in the Eastern North American-East Asian genus *Stewartia* (Theaceae: Stewartieae): insight from chloroplast and nuclear DNA sequence data. *Castanea* **67**, 290–301 (2002).
9. Lin, H.-Y. et al. Phylogenomic conflict resulting from ancient introgression following species diversification in *Stewartia* s.l. (Theaceae). *Mol. Phylog. Evol.* **135**, 1–11 (2019).
10. Spongberg, S. A. & Fordham, A. J. *Stewartias - Small Trees and Shrubs for All Seasons*. *Arnoldia* **35**, 165–180 (1975).
11. Tautz, D. & Domazet-Lošo, T. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12**, 692–702 (2011).
12. Wang, W. et al. High rate of chimeric gene origination by retroposition in plant genomes. *The Plant Cell* **18**, 1791–1802 (2006).
13. Li, Z.-W. et al. On the origin of de novo genes in Arabidopsis thaliana populations. *Genome Biol. Evol.* **8**, 2190–2202 (2016).
14. Zhang, L. et al. Rapid evolution of protein diversity by de novo origination in *Oryza*. *Nat. Ecol. Evol.* **3**, 679–690 (2019).
15. Zhang, Y. E., Vibrantovski, M. D., Landback, P., Marais, G. A. B. & Long, M. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLoS Biol.* **8**, e1000494 (2010).

16. Ruiz-Orera, J. et al. Origins of de novo genes in human and chimpanzee. *PLoS Genet* **11**, e1005721 (2016).
17. Ruiz-Orera, J., Verdaguier-Grau, P., Villanueva-Cañas, J. L., Messeguer, X. & Albà, M. M. Translation of neutrally evolving peptides provides a basis for de novo gene evolution. *Nat. Ecol. Evol.* **2**, 890–896 (2018).
18. Wu, D.-D., Irwin, D. M. & Zhang, Y.-P. De novo origin of human protein-coding genes. *PLoS Genet* **7**, e1002379 (2011).
19. Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A. & Begun, D. J. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc. Natl. Acad. Sci. USA* **103**, 9935–9939 (2006).
20. Rogers, R. L. & Hartl, D. L. Chimeric genes as a source of rapid evolution in *Drosophila melanogaster*. *Mol. Biol. Evol.* **29**, 517–529 (2011).
21. Vakirlis, N. et al. A molecular portrait of de novo genes in yeasts. *Mol. Biol. Evol.* **35**, 631–645 (2017).
22. Marcet-Houben, M. & Gabaldón, T. Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS Biol.* **13**, e1002220 (2015).
23. Marcet-Houben, M. & Gabaldón, T. Horizontal acquisition of toxic alkaloid synthesis in a clade of plant associated fungi. *Fungal Genet. Biol.* **86**, 71–80 (2016).
24. Carvunis, A.-R. et al. Proto-genes and de novo gene birth. *Nature* **487**, 370–374 (2012).
25. Long, M., VanKuren, N. W., Chen, S. & Vranoski, M. D. New gene evolution: little did we know. *Annu. Rev. Genet.* **47**, 307–333 (2013).
26. Cheng, L. et al. Transcriptome-based variations effectively untangling the intraspecific relationships and selection signals in Xinyang Maojian tea population. *Front. Plant Sci.* **14**, 1114284 (2023).
27. Guo, J. et al. Phylotranscriptomics in Cucurbitaceae reveal multiple whole-genome duplications and key morphological and molecular innovations. *Mol. Plant* **13**, 1117–1133 (2020).
28. Cheng, L. et al. Phylogenomics as an effective approach to untangle cross-species hybridization event: a case study in the family Nymphaeaceae. *Front. Genet.* **13**, 1031705 (2022).
29. Huang, W. et al. A well-supported nuclear phylogeny of Poaceae and implications for the evolution of C4 photosynthesis. *Mol. Plant* **15**, 755–777 (2022).
30. Zhang, L. et al. Phylotranscriptomics resolves the phylogeny of Pooideae and uncovers factors for their adaptive evolution. *Mol. Biol. Evol.* **39**, msac026 (2022).
31. Meng, Z. et al. Characterization of a *Saccharum spontaneum* with a basic chromosome number of $x = 10$ provides new insights on genome evolution in genus *Saccharum*. *Theor. Appl. Genet.* **133**, 187–199 (2020).
32. Huang, J. et al. Nuclear phylogeny and insights into whole-genome duplications and reproductive development of Solanaceae plants. *Plant Commun.* **4**, 100595 (2023).
33. Jiang, Y. et al. The genome assembly of *Chimonanthus praecox* var. concolor and comparative genomic analysis highlight the genetic basis underlying conserved and variable floral traits of wintersweet. *Ind. Crops Prod.* **206**, 117603 (2023).
34. Project, A. G. The *Amborella* genome and the evolution of flowering plants. *Science* **342**, 1241089 (2013).
35. Zhang, L. et al. The water lily genome and the early evolution of flowering plants. *Nature* **577**, 79–84 (2020).
36. Jin, G. et al. New genes interacted with recent whole-genome duplicates in the fast stem growth of Bamboos. *Mol. Biol. Evol.* **38**, 5752–5768 (2021).
37. Jiao, Y. et al. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97–100 (2011).
38. Qiao, X. et al. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biol.* **20**, 38 (2019).
39. Zhang, Q., et al. Transcriptome-based phylogeny and whole-genome duplication in Theaceae. *bioRxiv*, (2021).
40. Xiang, Y. et al. Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Mol. Biol. Evol.* **34**, 262–281 (2017).
41. Huang, C. H. et al. Multiple polyploidization events across asteraceae with two nested events in the early history revealed by nuclear phylogenomics. *Mol. Biol. Evol.* **33**, 2820–2835 (2016).
42. Huang, F. et al. Genome assembly of the Brassicaceae diploid *Orychophragmus violaceus* reveals complex whole-genome duplication and evolution of dihydroxy fatty acid metabolism. *Plant Commun.* **4**, 100432 (2023).
43. Zhao, Y. et al. 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 (2021).
44. Lee, S., Choi, S., Jeon, D., Kang, Y. & Kim, C. Evolutionary impact of whole genome duplication in Poaceae family. *J. Crop Sci. Biotechnol.* **23**, 413–425 (2020).
45. Huang, S. et al. Draft genome of the kiwifruit *Actinidia chinensis*. *Nat. Commun.* **4**, 2640 (2013).
46. Wang, Y. et al. An ancient whole-genome duplication event and its contribution to flavor compounds in the tea plant (*Camellia sinensis*). *Hort. Res.* **8**, 176 (2021).
47. Leebens-Mack, J. H. et al. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
48. Zhang, Q. J. et al. The chromosome-level reference genome of tea tree unveils recent bursts of non-autonomous LTR retrotransposons in driving genome size evolution. *Mol. Plant* **13**, 935–938 (2020).
49. Xia, E. et al. The reference genome of tea plant and resequencing of 81 diverse accessions provide insights into its genome evolution and adaptation. *Mol. Plant* **13**, 1013–1026 (2020).
50. Lin, P. et al. The genome of oil-*Camellia* and population genomics analysis provide insights into seed oil domestication. *Genome Biol.* **23**, 14 (2022).
51. Gong, W. et al. Chromosome-level genome of *Camellia lanceoleosa* provides a valuable resource for understanding genome evolution and self-incompatibility. *Plant J* **110**, 881–898 (2022).
52. Wang, X. et al. Population sequencing enhances understanding of tea plant evolution. *Nat. Commun.* **11**, 1–10 (2020).
53. Wei, C. et al. Draft genome sequence of *Camellia sinensis* var. *sinensis* provides insights into the evolution of the tea genome and tea quality. *Proc. Natl. Acad. Sci. USA* **115**, E4151–E4158 (2018).
54. Xia, E.-H. et al. The tea tree genome provides insights into tea flavor and independent evolution of caffeine biosynthesis. *Mol. Plant* **10**, 866–877 (2017).
55. Zhang, L. et al. Chemistry and Biological Activities of Processed *Camellia sinensis* Teas: A Comprehensive Review. *Compr. Rev. Food Sci. Food Saf.* **18**, 1474–1495 (2019).
56. Zeng, L., Watanabe, N. & Yang, Z. Understanding the biosyntheses and stress response mechanisms of aroma compounds in tea (*Camellia sinensis*) to safely and effectively improve tea aroma. *Crit. Rev. Food Sci. Nutr.* **59**, 2321–2334 (2019).
57. Yu, Z. & Yang, Z. Understanding different regulatory mechanisms of proteinaceous and non-proteinaceous amino acid formation in tea (*Camellia sinensis*) provides new insights into the safe and effective alteration of tea flavor and function. *Crit. Rev. Food Sci. Nutr.* **60**, 844–858 (2020).
58. Zuo, H. et al. Dynamic change of tea (*Camellia sinensis*) leaf cuticular wax in white tea processing for contribution to tea flavor formation. *Food Res. Int.* **163**, 112182 (2023).
59. Ho, C.-T., Zheng, X. & Li, S. Tea aroma formation. *Food Sci. Hum. Wellness* **4**, 9–27 (2015).

60. Xia, E.-H. et al. Tea plant genomics: achievements, challenges and perspectives. *Hort. Res.* **7**, 7 (2020).
61. Pichersky, E. & Lewinsohn, E. Convergent evolution in plant specialized metabolism. *Annu. Rev. Plant Biol.* **62**, 549–566 (2011).
62. Dunning Hotopp, J. C. Horizontal gene transfer between bacteria and animals. *Trends Genet.* **27**, 157–163 (2011).
63. Li, Y. et al. HGT is widespread in insects and contributes to male courtship in lepidopterans. *Cell* **185**, 2975–2987.e2910 (2022).
64. Gogarten, J. P., Fournier, G. & Zhaxybayeva, O. Gene transfer and the reconstruction of life's early history from genomic data. *Space Sci. Rev.* **135**, 115–131 (2008).
65. Jones, J. T., Furlanetto, C. & Kikuchi, T. Horizontal gene transfer from bacteria and fungi as a driving force in the evolution of plant parasitism in nematodes. *Nematology* **7**, 641–646 (2005).
66. Rice, D. W. et al. Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm *Amborella*. *Science* **342**, 1468–1473 (2013).
67. Tian, C. et al. A gene expression map of shoot domains reveals regulatory mechanisms. *Nat. Commun.* **10**, 141 (2019).
68. Hembree, W. G., Ranney, T. G., Jackson, B. E. & Weathington, M. J. H. Cytogenetics, ploidy, and genome sizes of *Camellia* and related genera. *HortScience* **54**, 1124–1142 (2019).
69. Horiuchi, K. & Oginuma, K. Karyomorphology of three species of *Stewartia*(Theaceae) in Japan. *Chromosome science* **5**, 79–82 (2001).
70. Zhang, C. et al. Asterid phylogenomics/phylotranscriptomics uncover morphological evolutionary histories and support phylogenetic placement for numerous whole-genome duplications. *Mol. Biol. Evol.* **37**, 3188–3210 (2020).
71. Zachos, J. C., Dickens, G. R. & Zeebe, R. E. An early Cenozoic perspective on greenhouse warming and carbon-cycle dynamics. *Nature* **451**, 279–283 (2008).
72. Zhou, S. Y., Chen, Y., M., Guo, C., C. & Qi, J. PhyloMCL: Accurate clustering of hierarchical orthogroups guided by phylogenetic relationship and inference of polyploidy events. *Methods Ecol. Evol.* **11**, 43–954 (2020).
73. Long, M. & Langley, C. H. Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. *Science* **260**, 91–95 (1993).
74. Li, J. et al. The nearly complete assembly of the *Cercis chinensis* genome and Fabaceae phylogenomic studies provide insights into new gene evolution. *Plant Commun.* **4**, 100422 (2023).
75. Chen, J. D. et al. The chromosome-scale genome reveals the evolution and diversification after the recent tetraploidization event in tea plant. *Hort. Res.* **7**, 63 (2020).
76. Zhang, W. et al. Genome assembly of wild tea tree DASZ reveals pedigree and selection history of tea varieties. *Nat. Commun.* **11**, 3719 (2020).
77. Zhang, X. et al. Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant *Camellia sinensis*. *Nat. Genet.* **53**, 1250–1259 (2021).
78. Shen, T. F. et al. The reference genome of *Camellia chekiangoleosa* provides insights into *Camellia* evolution and tea oil biosynthesis. *Hort. Res.* **9**, uhah083 (2022).
79. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
80. Zhang, Q. et al. Phylotranscriptomics of Theaceae: generic level relationships, reticulation and whole-genome duplication. *Ann. Bot.* **129**, 457–471 (2022).
81. Larson, D. A., Walker, J. F., Vargas, O. M. & Smith, S. A. A consensus phylogenomic approach highlights paleopolyploid and rapid radiation in the history of Ericales. *Am. J. Bot.* **107**, 773–789 (2020).
82. Yang, F.-S. et al. Chromosome-level genome assembly of a parent species of widely cultivated azaleas. *Nat. Commun.* **11**, 5269 (2020).
83. Jaillon, O. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
84. Wang, J.-P. et al. Two likely auto-tetraploidization events shaped kiwifruit genome and contributed to establishment of the Actinidiaceae family. *iScience* **7**, 230–240 (2018).
85. Han, X. et al. Two haplotype-resolved, gap-free genome assemblies for *Actinidia latifolia* and *Actinidia chinensis* shed light on the regulatory mechanisms of vitamin C and sucrose metabolism in kiwifruit. *Mol. Plant* **16**, 452–470 (2023).
86. Soltis, P. S. & Soltis, D. E. Ancient WGD events as drivers of key innovations in angiosperms. *Curr. Opin. Plant Biol.* **30**, 159–165 (2016).
87. Vanneste, K., Baele, G., Maere, S. & Van de Peer, Y. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. *Genome Res.* **24**, 1334–1347 (2014).
88. Fawcett, J. A., Maere, S. & Van de Peer, Y. Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. *Proc. Natl. Acad. Sci. USA* **106**, 5737–5742 (2009).
89. Long, M., Betrán, E., Thornton, K. & Wang, W. The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* **4**, 865–875 (2003).
90. Jander, G. & Barth, C. Tandem gene arrays: a challenge for functional genomics. *Trends Plant Sci.* **12**, 203–210 (2007).
91. Ma, X. et al. Seagrass genomes reveal ancient polyploidy and adaptations to the marine environment. *Nat. Plants* **10**, 240–255 (2024).
92. Liu, J. N. et al. Genomic analyses provide insights into the evolution and salinity adaptation of halophyte *Tamarix chinensis*. *GigaScience* **12**, giad053 (2023).
93. Qian, Z., Ding, J., Li, Z. & Chen, J. The high-quality *Pinellia pedatisecta* genome reveals a key role of tandem duplication in the expansion of its agglutinin genes. *Hort. Res.* **10**, uhac289 (2022).
94. Pahlich, E. & Gerlitz, C. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemistry* **19**, 11–13 (1980).
95. Wingett, S. et al. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research* **4**, 1310 (2015).
96. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
97. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
98. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
99. Ou, S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob. DNA* **10**, 48 (2019).
100. Ou, S. & Jiang, N. LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol* **176**, 1410–1422 (2018).
101. Zhang, Y., Park, C., Bennett, C., Thornton, M. & Kim, D. Rapid and accurate alignment of nucleotide conversion sequencing reads with HISAT-3N. *Genome Res* **31**, 1290–1295 (2021).
102. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
103. Shumate, A., Wong, B., Perte, G. & Perte, M. Improved transcriptome assembly using a hybrid of long and short reads with StringTie. *PLoS Comp. Biol.* **18**, e1009730 (2022).

104. Stanke, M., Tzvetkova, A. & Morgenstern, B. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol.* **7**, S11 (2006).
105. Delcher, A. L., Bratke, K. A., Powers, E. C. & Salzberg, S. L. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673–679 (2007).
106. Campbell, M. S., Holt, C., Moore, B. & Yandell, M. Genome annotation and curation using MAKER and MAKER-P. *Current protocols in bioinformatics* **48**, 1–39 (2014).
107. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
108. Paysan-Lafosse, T. et al. InterPro in 2022. *Nucleic Acids Res.* **51**, D418–D427 (2023).
109. Potter, S. C. et al. HMMER web server: 2018 update. *Nucleic Acids Res.* **46**, W200–W204 (2018).
110. Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. *Methods Mol. Biol.* **1962**, 1–14 (2019).
111. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
112. Mu, W. et al. The draft genome assembly of the critically endangered *Nyssa yunnanensis*, a plant species with extremely small populations endemic to Yunnan Province, China. *GigaByte (Hong Kong, China)* **2020**, 1–2 (2020).
113. Yang, X. et al. A chromosome-level genome assembly of the Chinese tupelo *Nyssa sinensis*. *Sci. Data* **6**, 282 (2019).
114. Chen, Y. et al. Genomic analyses of a “living fossil”: The endangered dove-tree. *Mol. Ecol. Resour.* **20**, 756–769 (2020).
115. Nashima, K. et al. Genome sequence of *Hydrangea macrophylla* and its application in analysis of the double flower phenotype. *DNA Res.* **28**, dsaa026 (2021).
116. Lamesch, P. et al. The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202–D1210 (2012).
117. Filiault, D. L. et al. The *Aquilegia* genome provides insight into adaptive radiation and reveals an extraordinarily polymorphic chromosome with a unique history. *Elife* **7**, e36426 (2018).
118. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
119. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
120. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
121. van Dongen, S. & Abreu-Goodger, C. Using MCL to extract clusters from networks. *Methods Mol. Biol.* **804**, 281–295 (2012).
122. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
123. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612 (2006).
124. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
125. Minh, B. et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
126. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermini, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
127. Zhang, C., Celine, S., Molloy, E. K. & Siavash, M. ASTRAL-Pro: quartet-based species-tree inference despite paralogy. *Mol. Biol. Evol.* **37**, 3292–3307 (2020).
128. Hochuli, P. A. & Feist-Burkhardt, S. Angiosperm-like pollen and *Afropollis* from the Middle Triassic (Anisian) of the Germanic Basin (Northern Switzerland). *Front Plant Sci.* **4**, 344 (2013).
129. Feng, X. et al. Genomic evidence for rediploidization and adaptive evolution following the whole-genome triplication. *Nat. Commun.* **15**, 1635 (2024).
130. Hughes N. F. *The enigma of angiosperm origins.* Cambridge University Press (1994).
131. Zeng, L. et al. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nat. Commun.* **5**, 1–12 (2014).
132. Doyle, J. A. & Hottón, C. L. Diversification of early angiosperm pollen in a cladistic context. *Pollen spores: patterns of diversification* **169**, 195 (1991).
133. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
134. Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L. L. & Hernández-Hernández, T. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytol* **207**, 437–453 (2015).
135. Chen, D., Zhang, T., Chen, Y., Ma, H. & Qi, J. Tree2GD: a phylogenomic method to detect large-scale gene duplication events. *Bioinformatics* **38**, 5317–5321 (2022).
136. Sun, P. et al. WGDl: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Mol. Plant* **15**, 1841–1851 (2022).
137. Wang, J. et al. A common whole-genome paleotetraploidization in Cucurbitales. *Plant Physiol* **190**, 2430–2448 (2022).
138. Chen, C. et al. TBtools-II: A “One for All, All for One” bioinformatics platform for biological big-data mining. *Mol. Plant* **13**, 1194–1202 (2023).
139. Mistry, J. et al. Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
140. Drost, H. G., Gabel, A., Liu, J., Quint, M. & Grosse, I. myTAL: evolutionary transcriptomics with R. *Bioinformatics* **34**, 1589–1590 (2018).
141. Domazet-Lošo, T. & Tautz, D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* **468**, 815–818 (2010).
142. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Brief. Bioinformatics* **18**, 205–214 (2017).
143. Zhang, Z. KaKs-Calculator 3.0: Calculating selective pressure on coding and non-coding sequences. *Genomics Proteomics Bioinformatics* **20**, 536–540 (2022).

Acknowledgements

We thank to Dr. Xinxin Zhu for providing the image materials of *S. sinensis*. We are particularly thankful for the anonymous reviewers’ valuable comments on the manuscript. We are grateful to the State Key Laboratory of Big Data, Guizhou University for providing super computing services for data analysis. This work was funded by the Program for Science & Technology Innovation Talents in Universities of Henan Province (HASTIT, No. 21HASTIT040) and the Scientific and Technological Project in Henan Province (232102521008). Additional support was provided by the Key Laboratory of Functional Agriculture in Higher Education of Guizhou Province, China, designated by the grant number Qian Jiao Ji [2023]007.

Author contributions

Yiyong Zhao, Hao Li, and Lin Cheng conceived and designed the research and experiments. Yiyong Zhao, Lin Cheng, Qunwei Han, Zhihan Zhang, and Yanlin Hao prepared samples and subsequent sequencing. Lin Cheng, Qunwei Han and Yanlin Hao performed most of the bioinformatic analyses under the assistance from Yiyong Zhao. Zhen Qiao, Mengge Li, Daliang Liu, Hao Yin, Tao Li and Ya Gao participated in analyzing the data. Wen Long provided support for the Linux cluster server. Shanshan Luo,

Houlin Yu and Xinhao Sun provided support for the data curation and storage. Critical revisions were made by Yiyong Zhao, Lin Cheng, Qunwei Han, Houlin Yu and Xinhao Sun. All authors thoroughly reviewed and endorsed the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42003-025-07525-x>.

Correspondence and requests for materials should be addressed to Hao Li or Yiyong Zhao.

Peer review information *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: David Favero and Mengtan Xing.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025