Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Integrative multi-omics analysis traces the divergent evolution of Chinese and Vietnamese Cinnamon and elucidates the biosynthetic pathway of cinnamaldehyde derivatives

Liqun Lin
Chinese Academy of Agricultural Sciences
Yuanxia Chen
Chinese Academy of Agricultural Sciences
Yingying Chen
Guangxi Forestry Research Institute
Shichao Sun
Chinese Academy of Agricultural Sciences
Xiaojing Liang
Guangxi Forestry Research Institute
Yongjin Zhu
Guangxi Forestry Research Institute
Chansan Zhu
Guangxi Forestry Research Institute
Siyu Wu
Guangxi Forestry Research Institute
Xiangyan Zeng
Guangxi Forestry Research Institute
Rihong Jiang
Guangxi Forestry Research Institute
Kaixiang Li
Guangxi Forestry Research Institute
Li Wang
wangli03@caas.cn

Chinese Academy of Agricultural Sciences

Research Article

Keywords: divergent selection, magnolia, 2-methoxycinnamaldehyde, medicinal plants, O-methyltransferase (OMT), phenylpropanoid biosynthesis

Posted Date: March 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8704689/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Human interaction with plants reshapes genetic diversity across species and lineages, yet this process remains poorly understood in perennial woody plants cultivated for non-fruit products. Cinnamon, primarily derived from *Cinnamomum cassia*, represents a compelling subject to investigate how ecological adaptation and long-term cultivation have shaped its genomic diversity and driven variation in its key bioactive compounds, namely cinnamaldehyde and its derivatives.

Results

To investigate the divergence of Chinese cinnamon (*C. cassia* var. *cassia*) and Vietnamese cinnamon (*C. cassia* var. *macrophyllum*), we integrated a *de novo* genome assembly, whole-genome resequencing of 98 accessions, and metabolomic and transcriptomic profiling of multiple tissues of both varieties. Phylogenomic and demographic analyses revealed that the two varieties diverged prior to documented human utilization. Vietnamese cinnamon formed a basal lineage with high genetic diversity, whereas Chinese cinnamon exhibited reduced genetic diversity and weak population structure, likely resulting from prolonged intensive cultivation. Genome-wide selection scans indicated genes involved in phenylpropanoid biosynthesis were under divergent selection. Metabolomic and transcriptomic profiling uncovered differentiation in cinnamaldehyde derivatives across tissues and varieties and pinpointed candidate enzymes in their biosynthesis. Functional validation demonstrated that two O-methyltransferases, Ccas10G000326 and Ccas05G001459, effectively enhance the transformation of 2-hydroxycinnamaldehyde into 2-methoxycinnamaldehyde.

Conclusions

Our findings demonstrate that persistent cultivation of a limited gene pool diminishes genetic diversity in perennial trees, whereas lineages under lower utilization pressure preserve greater evolutionary potential. It highlights that conserving and integrating underutilized genetic resources is essential for the sustainable use and improvement of woody crops.

Background

Different species or intraspecific lineages (such as subspecies and varieties) have been selectively utilized as economic crops to meet diverse regional needs under specific cultural and environmental contexts [1]. Although domestication is generally expected to reduce genetic diversity through the fixation of favorable alleles and bottlenecks, different taxonomic groups employed for similar agricultural ends may bear divergent genetic footprints, driven by their distinct cultivation practices and histories [2]. A well-known example is Asian rice (*Oryza sativa*), whose *indica* and *japonica* subspecies together

underpin global rice production [3]. While both were domesticated from a common ancestor, *O. sativa* ssp. *japonica* spread to higher latitudes and altitudes, experiencing strong founder effects that significantly reduced its effective population size and genetic diversity [4]. In contrast, *O. sativa* ssp. *indica* was primarily cultivated in its native tropical lowlands and thus avoided an equally severe genetic bottleneck; occasional gene flow from local wild or weedy rice populations further contributed to its relatively high genetic diversity [2, 5].

Such studies have been strongly biased toward annual grain crops [6–8] and, to a lesser extent, perennial fruit trees [9, 10]. By contrast, perennial woody plants cultivated primarily for non-fruit products, such as spices, medicinal materials, and resins, have received far less attention [11]. In these systems, where harvested traits are often chemically complex and not directly visible, the genetic consequences of long-term selective use and the interaction among ecological differentiation, chemical diversity, and human utilization remain largely unexplored.

Cinnamomum cassia (Lauraceae) represents a particularly compelling subject to fill in this gap, as Chinese (*C. cassia* var. *cassia*) and Vietnamese (*C. cassia* var. *macrophyllum*) cinnamon were widely cultivated in tropical and subtropical Asia [12] and the mountainous regions of northern and central Vietnam, respectively. The latter exhibited markedly large leaves, approximately twice as large as that of Chinese cinnamon (Fig. 1A), fast growth, and high cinnamaldehyde content [13–15], indicating a distinct metabolic profile shaped by local adaptation. Although Vietnamese cinnamon was historically treated as a separate species (e.g., *C. loureirii*), recent assessments based on morphological evidence support its treatment as a variety of *C. cassia* [16]. However, related genetic evidence has been limited [13]. This uncertainty is further compounded by overlapped distributions in northern Vietnam and southern China of the two varieties, shared essential oil profiles [17, 18], and the widespread introduction of Vietnamese germplasm into southern China through cultivation [14, 19, 20]. Furthermore, both Chinese and Vietnamese cinnamon are recognized as authorized sources in China Pharmacopeia (2025), implying equivalent biochemical and pharmacological values of the two varieties. Thus, key questions remain unresolved: the extent to which these two varieties represent ecologically differentiated lineages, how they are genetically and chemically related, and how human utilization has shaped their divergence.

Addressing these questions is of both scientific and practical importance, given the global prominence of cinnamon. Renowned for its enchanting aromatic allure and therapeutic benefits, cinnamon (*Cinnamomum spp.*) is among the most widely used spices worldwide, and has played a prominent role in culinary practices and traditional herb medicine for millennia [14, 21]. Global demand for cinnamon continues to increase [21],with *C. cassia* accounting for over 80% of the international trade [22]. *C. cassia* is revered as a remarkably versatile plant, with each tissue utilized for differentiated therapeutic purpose in traditional Chinese medicine (TCM). However, the biochemical basis underlying the differential therapeutic functions of each tissue remain poorly understood.

C. cassia contains over 160 bioactive compounds [23], among which *trans*-cinnamaldehyde (TCA) and its derivatives, such as 2-methoxycinnamaldehyde (2-MCA) and 2-hydroxycinnamaldehyde (2-HCA), are

the principal bioactive constituent, exhibiting potent anti-cancer and anti-inflammatory effects [24, 25]. TCA is biosynthesized via the phenylpropanoid pathway, originating from phenylalanine via cinnamic acid and cinnamoyl-CoA intermediates (Fig. 5B) [26, 27]. While this upstream pathway is relatively well characterized, the downstream biosynthetic steps leading to its key derivatives remain poorly understood. Elucidating the downstream steps is essential not only for understanding tissue-specific and inter-varietal metabolic differentiation, but also for enabling future metabolic engineering and targeted breeding efforts aimed at enhancing valuable pharmacological compounds [28].

Here, based on one *de novo* assembled genome of *C. cassia* var. *cassia*, whole genome resequencing data of 98 accessions, as well as metabolomic and transcriptomic profiling of key tissues of both Chinese and Vietnamese cinnamon, we aim to investigate the genetic diversity of *C. cassia*, with particular focus on genetic differentiation between Chinese and Vietnamese cinnamon, to dissect the biochemical basis of distinct pharmacological function of different tissues of cinnamon and biochemical differentiation between the two varieties, and to identify the key enzymes responsible for the production of major cinnamaldehyde derivatives. The study is poised to offer novel insights into evolutionary history of cinnamon, hold significant promise for germplasm conservation and proper pharmacological usage, provide novel breeding strategies to enhance the content of anti-cancer compounds, and standardize the international trade of cinnamon.

Results

Genome assembly and gene annotation

The k-mer analysis (K = 21) revealed that the genome size of *C. cassia* was approximately 1.18 Gb, with a heterozygosity rate of 0.51% and 57.59% repetitive sequences (Fig. S1). 62.07 Gb PacBio circular consensus sequencing (CCS) long reads (55.42× coverage) with an average length of 14.72 kb and 140.45 Gb high-throughput chromosome conformation capture (Hi-C) data (125.40× coverage) were generated for genome assembly of *C. cassia* (Table S1). The resulting genome was assembled into 12 pseudochromosomes and 751 unanchored scaffolds, with a total length of 1.12 Gb and scaffold N50 of 91.36 Mb (Fig. 1B). The high fidelity of the assembly was verified by three key metrics. First, the mapping rates of RNA-seq data to the assembled genome were as high as 87.62–95.77%. Second, Benchmarking Universal Single-Copy Orthologs assessment (BUSCO) analysis showed that a total of 1572 complete genes (97.4% among the conserved genes in “embryophyte_odb10” database) were detected in the assembled genome, suggesting high completeness at the gene level. Third, the long terminal repeat (LTR) assembly index (LAI) score [29] was 17.26, which is classified as “reference” quality, reflecting a highly complete assembly of transposable elements (TEs) (Table S2).

Based on multiple-tissue RNA-Seq data and *ab initio* prediction using the deep learning tool Helixer (Holst et al., 2023), a total of 38,639 protein-coding genes were identified with an average gene length of 7,466.09 bp. Approximately 84.32% of the protein-coding genes were functionally annotated in the Pfam,

GO and KEGG databases. The annotated repeated sequences accounted for 59.51% of the *C. cassia* genome, including 41.31% of LTRs, 10.31% Copia, and 25.5% Gypsy (Table S2).

Phylogenetic and Whole Genome Duplication (WGD) Analyses

The phylogenetic tree was constructed based on the concatenated 337 single-copy orthologous genes among 14 species. *C. cassia* was clustered with other *Cinnamomum* species in the magnoliids lineage. It supported the sister relationship between magnoliids and eudicots, which was consistent with some previous studies (Fig. 2A) [30–32]. The MCMC time tree with fossil calibrations indicated that *C. cassia* diverged from *Cinnamomum burmanni* at approximately 7.5 million years ago (MYA). The genus *Cinnamomum* and *Camphora* diverged at about 13.9 MYA. Consistent with previous studies (Li et al., 2022), we detected two shared WGD events in Lauraceae (*C. cassia*, *C. burmanni* and *Camphora kanehirae*), represented by two signature peaks where the synonymous substitution rates (Ks) were approximately 0.6 and 0.8 (Fig. 2B and C). And we did not detect any *C. cassia*-specific WGD event.

We subsequently identified 887 contracted ($P < 0.005$) and 1074 expanded ($P < 0.005$) gene families in *C. cassia.* Further GO and KEGG enrichment analysis of expanded genes revealed significant enrichment in genes involved in secondary metabolite biosynthetic process, such as phenylpropanoid biosynthesis, terpenoid biosynthetic process, and flavonoid biosynthesis (Fig. 2D and E).

Population genetic structure and demographic history of C. cassia

Whole-genome resequencing of 98 individuals was aligned to the reference genome of *C. cassia*, generating 60,259,942 biallelic SNPs, with an average mapping depth of 22.10×. We first excluded individuals with ambiguous taxonomic assignment or excessive missing SNPs (missing rate > 20%), as well as duplicated or closely related samples (kinship coefficient > 0.354, corresponding to first-degree relatives or identical individuals). In addition, SNPs, which were missing among > 50% individuals, were removed (Fig. S2). This initial filtering step resulted in dataset D1, comprising 88 individuals (with one *C. heyniaum* individual as outgroup) and 59,910,409 SNPs (Table S3). To obtain a more stringently curated dataset for downstream population genomic analyses, we further removed second cousins or more closely related individuals (kinship coefficient > 0.044), excluded admixed individuals with the major ancestry coefficients < 0.7 inferred from ADMIXTURE analyses, and filtered out low-frequency variants (minor allele frequency < 0.01). Applying these filters yielded the final high-confidence dataset D2, comprising 43 individuals and 867,268 SNPs.

To explore the relationship of these accessions, dataset D1 were used to perform population structure analysis, phylogenetic analysis and principal component analysis (PCA). According to the lowest cross-validation (CV) error in ADMIXTURE analysis, the number of optimal ancestral components was 2 (Fig. 3A). At $K = 2$, the 87 *C. cassia* accessions were divided into two major genetic lineages, the first lineage comprises *C. cassia* var. *macrophylla* (hereafter “macro”), predominantly introduced from Vietnam; the second lineage includes *C. cassia* var. *cassia* (hereafter “cassia”), cultivated across four major growing

regions in China. The accessions within the cassia lineage were divided into two subgroups: cassia-FCG (membership coefficients < 0.90, the earliest-diverging lineage within cassia), mainly collected from a coastal and border city adjacent to Vietnam, Fangchenggang; cassia-MDA (membership coefficients > 0.90), representing accessions from the main distributional areas in China (Fig. 3A and B). PCA analysis showed consistent clustering of the samples with the ADMIXTURE and Neighbor-joining (NJ) tree results, with the first two components explaining 13.3% and 6.6% of the total variance (Fig. 3C).

The cassia-FCG subgroup showed signs of genetic admixture. However, treeMix, f3/f4 statistics, and Dsuite analyses based on samples selected from the D2 dataset did not detect statistically significant gene flow between macro and cassia subgroups, suggesting that admixture—if present—is either ancient or limited in scope (Fig. S3, Table S4). Conversely, cassia-MDA samples exhibited weak geographic clustering, consistent with extensive human-mediated movement of propagules, vegetative propagation, and regional exchange of elite cultivars, consistent with previous reports [33].

Nucleotide diversity (π) and population differentiation levels (Pairwise fixation index; F_{st}) were also estimated for these identified groups. Macro exhibited the highest genetic diversity (π = 0.00485), followed by cassia-FCG (π = 0.00374) and cassia-MDA (π = 0.00339) (Fig. 3D). Substantial genetic differentiation was observed between macro and cassia-MDA (F_{st} = 0.253), with reduced divergence between Macro and cassia-FCG (F_{st} = 0.188). In contrast, cassia-FCG and cassia-MDA exhibited minimal genetic differentiation (F_{st} = 0.084), consistent with their shared ancestry (Fig. 3D). Furthermore, the sliding windows with the top 1% high F_{st} values between macro and cassia-MDA were considered under divergent selection, yielding 412 windows encompassing or proximal to 381 annotated genes (Fig. S4). Functional enrichment analyses of these genes revealed a significant enrichment in phenylpropanoid biosynthesis and related secondary metabolic processes, with enriched molecular functions dominated by O-methyltransferase-related activities (Fig. 3E). Notably, 21 out of the 22 genes associated with O-methyltransferase activity were located on chromosome 5, within the region from 71.1 to 75.4 Mb. Consistently, a local Manhattan plot of F_{st} values revealed pronounced peaks on chromosome 5 (Fig. 3F), where multiple adjacent high-F_{st} windows clustered into a continuous block. Within this region, more than ten OMT genes were located within or immediately flanking the selected windows (Fig. 3F), indicating that the chromosome 5 OMT cluster represents a major genomic hotspot of population differentiation between the macro and cassia-MDA groups.

Demographic history inferred with PSMC revealed a population expansion in both macro and cassia-MDA, followed by a gradual decline at around 30 MYA (Fig. S5). To enhance resolution for recent history, SMC + + analysis was performed using SNP data. SMC + + analysis corroborated the demographic trends revealed in PSMC and estimated the divergence between macro and cassia-MDA at ~ 400,000 years ago, and the split between cassia-MDA and cassia-FCG at ~ 78,500 years ago (Fig. 3G).

Volatile metabolic differentiation among tissues provides a chemical basis for tissue-specific medicinal applications of C. cassia

The overall volatile compound composition of Chinese cinnamon (using bark, fruit, leaf, petiole, and twig samples) and Vietnamese cinnamon (using bark, leaf, and twig samples) were initially compared. Despite differences in sampled tissues, the two varieties displayed broadly similar metabolic composition, with terpenoids as the predominant class (28.0% and 25.1%, respectively) and only modest shifts in the relative proportions of esters, heterocyclic compounds, and ketones. And we note that aldehydes, including TCA, 2-HCA and 2-MCA, were present at comparable levels between varieties (6.1% vs. 7.0%), ranking as the sixth most abundant class (Fig. 4A). A greater number of volatile metabolites were detected in Vietnamese cinnamon than in Chinese cinnamon (1236 vs. 686). However, this difference should be interpreted with caution, as the datasets were generated nearly two years apart using expanded spectral libraries, which may have increased detection sensitivity. Although previous studies have also reported a richer chemical repertoire in Vietnamese cinnamon [16], suggesting that genuine biological differences may exist, parallel analyses under identical experimental conditions will be required to disentangle technical effects from biological variation.

Based on the untargeted volatile compounds of five tissues in Chinese cinnamon, we next examined tissue-specific divergence in volatile metabolite composition. PCA revealed strong metabolic differentiation among tissues: the first two principal components accounted for 43.1% and 19.7% of the total variance, respectively, with PC1 clearly separating the five tissues (Fig. 4B). Notably, leaves exhibited distinct biochemical compounds compared to the other tissues, whereas petioles and twigs showed similar metabolic profiles, likely due to their tissue homology.

Tissue-specific signatures were further evaluated with Venn diagrams. Among the 686 volatile metabolites detected in five tissues of Chinese cinnamon, approximately 40 had relative contents exceeding 1%, and these were consistently present across all five tissues (Table S5). With the threshold gradually increased, from the top 1% to the top 30, 25, 20, 15, and 10 metabolites, it revealed a clear shift from commonly shared metabolites among all tissues to tissue-specific profiles (Fig. 4C). It suggested that while all five tissues shared a fundamental volatile metabolic diversity, their dominant metabolite compositions are tissue dependent. The specialized metabolic patterns likely contribute to both the unique and shared pharmacological characteristics observed across tissue types. Bark exhibited the highest degree of metabolic specialization among all tissues. Five of its ten most abundant compounds were found to be exclusive to bark and were absent from the top-ten lists of other tissues (Fig. 4C). Notably, *α*-copaene was the most enriched constituent in bark, in contrast to the cinnamaldehyde-dominated profiles observed in other tissues (Table S5). This bioactive sesquiterpene, also found in several aromatic and historically important medicinal plants such as myrrh (*Commiphora myrrha*), has been reported to possess strong antibacterial activity against common pathogens, with minimum inhibitory concentrations comparable to those of cinnamaldehyde [34–36]. Likely, fruits, the only tissue documented for use in treating cold-related abdominal discomfort in TCM, possibly correlated with its high *β*-guaiene content, which is consistent with other *β*-guaiene-rich medicinal plants across different ethnomedical traditions [37]. Together, these findings suggest that tissue-specific enrichment of key bioactive volatiles underpins both the shared and differentiated medicinal roles of *C. cassia* tissues,

providing a chemical rationale for their distinct traditional uses, not that any function is exclusive to a single tissue.

Functional OMT genes involving in biosynthesis of 2-MCA

To investigate the genetic basis underlying tissue- and variety-dependent differences in cinnamaldehyde derivatives, we first quantified three major cinnamaldehyde-related compounds (TCA, 2-MCA, and 2-HCA) across multiple tissues using LC-MS under a unified analytical batch (Table S6). In the three tissues commonly sampled from both varieties (leaf, bark, and twig), concentrations of all three metabolites were significantly higher in Chinese cinnamon than in Vietnamese cinnamon. Despite the difference in absolute abundance, both varieties displayed highly consistent tissue-dependent distribution patterns (Fig. 5A). Specifically, TCA accumulated predominantly in twigs (16.8 μg/mg in Chinese cinnamon; 6.3 μg/mg in Vietnamese cinnamon), followed by bark (12.1 and 5.7 μg/mg, respectively), and was the lowest in leaves (8.5 and 1.6 μg/mg). In contrast, both 2-HCA and 2-MCA were the most abundant in leaves and the lowest in twigs (Fig. 5A). Notably, the most pronounced inter-varietal difference was observed for 2-MCA: 2-MCA levels in the leaf of Chinese cinnamon reached approximately 12.0 μg/mg, more than ten-fold higher than that in Vietnamese cinnamon (around 0.8 μg/mg; Welch's *t*-test, $P < 0.001$). In Chinese cinnamon, we further quantified metabolites in two additional tissues—young bark and petioles—which were not available for Vietnamese cinnamon (Fig. S6). TCA abundantly accumulated in petioles (25.1 μg/mg) and in young bark (16.6 μg/mg), whereas both 2-HCA and 2-MCA showed relatively low levels in these tissues compared with leaves. Across all tissues examined, 2-HCA was consistently detected at low levels across all tissues (0.005-2.9 μg/mg), with a distribution pattern largely paralleling that of 2-MCA, suggesting that it may represent a low-abundance intermediate in downstream modification of TCA (Fig. S6). It should be noted that, although samples of the two varieties were collected from the same location, tree age and developmental stag

Based on these quantitative patterns, we next sought to elucidate the downstream biosynthetic step responsible for the formation of 2-MCA. We hypothesized that TCA undergoes hydroxylation to yield 2-HCA, followed by O-methylation to produce 2-MCA (Fig. 5B). However, the consistently low abundance of 2-HCA in Chinese cinnamon, together with the large and functionally diverse gene families potentially involved in hydroxylation, rendered direct identification of the upstream oxidase challenging. Furthermore, given that OMT-related functional categories were strongly enriched among regions showing high population differentiation, we prioritized a systematic investigation of OMT genes as an essential step toward reconstructing the downstream biosynthetic pathway.

Using an hmmsearch-based approach, we identified 88 putative OMT genes (> 180 amino acids) in the *C. cassia* genome. And these genes showed a highly non-random chromosomal distribution, with 40 genes (45.45%) located on chromosome 5 (Fig. S7). Phylogenetic analysis incorporating 52 experimentally verified plant OMTs (Table S7) demonstrated that OMTs cluster predominantly by substrate preference rather than species origin and that OMTs of *C. cassia* were classified into 10 CCoAOMTs and 78 COMTs (Fig. 5C). Based on the phylogeny, we excluded genes located in clades associated with

benzylisoquinoline/alkaloid metabolism, flavonoid and isoflavonoid pathways, phenylpropene and coumarin biosynthesis, carboxylic acid methylation, and CCoA-related OMTs. Instead, we focused on OMTs within or proximal to the hydroxycinnamaldehyde / hydroxycinnamic acid / hydroxycinnamyl alcohol (CAL) clade, generating a refined OMT candidate set. Integration of transcriptomic data further narrowed this set to 21 highly expressed OMT genes, three of which were overlapped with the selected OMT genes on chromosome 5 (Fig. 3F). Functional assays using an *Escherichia coli* transient expression system demonstrated that two genes, Ccas10G000326, and Ccas05G001459, effectively catalyzed the C-2 hydroxyl-to-methoxy conversion, significantly enhancing the conversion of 2-HCA into 2-MCA (Fig. 5D). Among these, Ccas05G001459 was located in the vicinity of a high-F_{st} window identified in genome-wide selection scan, whereas Ccas10G000326 was not associated with any detected differentiation outlier. Detailed inspection of sequence variation surrounding Ccas05G001459 revealed that SNPs within the coding region exhibited low F_{st} values, while the highest F_{st} signals were detected approximately 20 kb upstream of the gene.

Discussion

This study demonstrates that Chinese and Vietnamese cinnamon, two naturally diverged varieties of *C. cassia*, exhibited genetic differentiation arising from ecological adaptation, resulting in distinct accumulation patterns of phenylpropanoid metabolites. Differences in subsequent cultivation practices further reshaped their genetic trajectories: prolonged and intensive exploitation of Chinese cinnamon resulted in pronounced genetic reduction, whereas the short-term and low-demand use of Vietnamese cinnamon enabled it to retain substantially higher genetic diversity. Altogether, this contrast illustrates how human utilization can either erode or preserve standing genetic variation in perennial economic plants. Our work provides a framework for understanding how ecological adaptation and human utilization shape genetic and biochemical diversities in understudied woody crops, with direct implications for taxonomic classification, conservation, and sustainable breeding.

By leveraging genome-wide SNPs derived from approximately 59 million SNPs across more than 90 accessions, this study provides the first comprehensive genomic framework for evaluating the taxonomic relationship between Chinese and Vietnamese cinnamon. Phylogenomic analyses show that the Vietnamese cinnamon lineage (macro group) diverged first from the common ancestor and occupied a basal position within all *C. cassia* accessions (Fig. 3B), rejecting the view that Vietnamese cinnamon merely represents a processed form of Chinese cinnamon. Additionally, the two lineages are not reciprocally monophyletic (Fig. 3B), indicating that the degree of divergence does not justify species-level separation. These findings are congruent with, but more definitive than previous ISSR-based studies (Zheng, Xu [13], strongly support recognizing Vietnamese cinnamon as a large-leaf variety of *C. cassia* rather than as an independent species.

Demographic inference further indicates that the divergence between the two varieties occurred approximately 0.4 MYA (Fig. 3G), well before any documented human utilization of cinnamon [12]. It supports a natural divergence, rather than that Chinese cinnamon domesticated from Vietnamese

cinnamon by anthropogenic selection. Although geographically proximate, the occurrence of the two varieties is nearly separated via the biogeographical boundary between tropical and subtropical floristic regions in southern China and mainland of Southeast Asia, where strong ecological gradients are common [38]. Ecological adaptation of the two varieties might lead to the pronounced differences in leaf size and other traits [39, 40]. Genome-wide scans for signals of divergent selection revealed that regions exhibiting extremely high F_{st} values between macro and cassia were significantly enriched for genes involved in phenylpropanoid biosynthesis and related secondary metabolic processes (Fig. 3E). These pathways, including flavonoid, lignin, and pigment biosynthesis as well as multiple O-methyltransferase functions, are well known to mediate biotic and abiotic stress tolerance [41, 42]. Notably, only one functionally validated OMT gene, Ccas05G001459, was located near a highly differentiated region, yet no significant coding sequence divergence or expression difference of this gene was detected between the two varieties. The genomic region exhibiting the strongest differentiation was instead located in upstream regions of OMT genes, indicating that selection acted on regulatory elements or the region may be in a linkage group with nearby selected loci. In addition, our enzyme assays focused specifically on differences in phenylpropanoid metabolism between the two varieties — particularly cinnamaldehyde derivatives — the overall metabolic divergence between Chinese and Vietnamese cinnamon is likely to be broader. It is possible that the selected OMT genes, though not directly implicated in the tested substrates, may be effective in catalyzing other substrates, including flavonoid and phenolic compound modification.

Despite extensive sampling across the major production regions of China, Chinese cinnamon exhibits unexpectedly low genetic diversity and weak population structure (Fig. 3A, D), with many locally recognized landraces lacking clear genetic differentiation (Fig. 3B; Table S3). This pronounced genetic homogenization contrasts sharply with the markedly higher diversity retained in Vietnamese cinnamon (Fig. 3D), despite of limited sampling. The contrasting genetic diversity of the two varieties likely reflects differences in the duration, intensity, and mode of utilization. Chinese cinnamon has been used in China for medicinal and culinary purposes for millennia [12], with sustained high demand driving early domestication, large-scale cultivation, and repeated propagation from a restricted genetic base. These practices are expected to reduce effective population size, impose strong demographic bottlenecks, and accelerate genetic erosion over time [43]. In contrast, although Vietnamese cinnamon has been utilized locally for century, its large-scale cultivation started only in recent decades, and this production continues to be characterized by low-intensity, heterogeneous management within smallholder agroforestry systems. In these systems, planting material is typically sourced from multiple local populations and natural regeneration is common [14], thereby avoiding strong directional selection and genetic fixation [44]. Consequently, Vietnamese cinnamon has likely escaped the genetic erosion associated with prolonged cultivation and now represents a comparatively intact genetic reservoir within the species.

The genetic erosion observed in Chinese cinnamon is not a single case but reflects a recurrent risk in under-managed perennial economic plants. Comparable patterns have been documented in rubber [45], olive [46] and banana [47], where intensive cultivation, clonal propagation, or reliance on elite lineages

has resulted in genetic homogenization and increased vulnerability to abiotic and biotic stresses. Unlike annual crops, where rapid generational turnover and active breeding facilitate diversification, perennial trees are characterized by long generation times, frequent clonal propagation, and extended cultivation cycles, which together slow the introduction of new variation and amplify the cumulative effects of genetic bottlenecks. Under such conditions, genetic erosion can accumulate gradually and often unnoticed, particularly when the primary traits under selection are chemically or anatomically cryptic rather than visually apparent. Consequently, for perennial trees, human use does not inevitably lead to domestication or genetic improvement; instead, prolonged exploitation of a narrow genetic base may progressively erode standing variation and constrain future adaptation.

This contrast between Chinese and Vietnamese cinnamon not only clarifies the genetic erosion in cultivated cinnamon, but also points directly toward several complementary strategies can be adopted to mitigate further genetic erosion in cinnamon and other perennial economic trees. First, systematic *in situ* and *ex situ* conservation of genetic resources is essential to safeguard remaining diversity, as demonstrated in olive [46], shea [48], and other long-lived tree crops [49, 50]. Comprehensive genetic surveys should be conducted to identify priority conservation areas, establish germplasm repositories, and characterize genetic diversity of locally recognized landraces to detect rare or residual alleles [50]. Second, breeding strategies for perennial crops can draw valuable lessons from annual crop improvement. Rather than pursuing *de novo* domestication, incorporating naturally differentiated yet cross-compatible lineages into improvement programs represents a low-risk, high-reward pathway to broaden the genetic base [51]. Analogous to the exploitation of indica–japonica diversity in rice [52], the integration of Vietnamese cinnamon—with its higher genetic diversity and favorable traits—into cinnamon breeding could enhance resilience and adaptive potential. Together, these approaches provide a realistic framework for balancing utilization, conservation, and genetic improvement in cinnamon and other understudied woody crops.

Conclusions

This study provides an integrated genomic, metabolomic, and transcriptomic framework for understanding the evolution, utilization, and biochemical diversification of *C. cassia*. Our results demonstrate that Chinese and Vietnamese cinnamon represent naturally diverged, ecologically differentiated lineages within *C. cassia*, rather than products of recent human-mediated differentiation. Long-term, intensive use of Chinese cinnamon has resulted in pronounced genetic homogenization, whereas the mild utilization of Vietnamese lineage retains substantially higher genetic diversity. By resolving both the evolutionary history and metabolic differentiation of these two varieties, this work highlights how selective utilization of intraspecific lineages can reshape genetic diversity in perennial trees. More broadly, our findings underscore the importance of conserving underutilized lineages as reservoirs of evolutionary potential and integrating them into future improvement programs to ensure the sustainable use of woody crops.

Methods

Plant materials

Five Chinese cinnamon trees and one Vietnamese cinnamon cultivated in the Resource Garden of Guangxi Academy of Forestry Sciences, Guangxi, China (22°33'13"N, 108°12'40"E) were selected for genomic sequencing, RNA sequencing (RNA-seq), and gas chromatography-mass spectrometry (GC-MS) analyses. All plants were grown under uniform conditions. Genomic DNA for *de novo* assembly was obtained from fresh leaves of a single Chinese cinnamon tree (*C. cassia* var. *cassia*; CC135). For RNA-seq analysis of Chinese cinnamon, five tissue types—bark, young bark, fruit, leaf, and petiole—were collected. For GC-MS analysis, a similar set of tissues was used, with twigs substituted for young bark due to differences in tissue availability and analytical suitability. All Chinese cinnamon tissues were collected from the same five individuals, with three biological replicates per tissue. For Vietnamese cinnamon, three tissue types—bark, leaf, and twig—were sampled for both RNA-seq and GC-MS analyses, each with three replicates.

Healthy leaves from 96 *C. cassia* accessions were collected from its main distribution area in China. Among them, 77 samples were collected in Guangxi, 12 in Guangdong, 4 in Yunnan, 2 in Fujian, and 1 in Hainan. Further, a *C. heyneaum* and a *Cinnamomum verum* were collected to use as outgroups.

Genomic DNA preparation, RNA extraction, and cDNA preparation

High–molecular weight (HMW) genomic DNA was extracted from fresh leaves of *C. cassia* var. *cassia* by the DNeasy Plant Mini Kit (Qiagen, United States). More than 50 μg of the HMW DNA was used to generate SMRTbell libraries. Long read sequencing was performed on the PacBio Sequel II platform with “circular consensus sequencing (CCS)” mode.

Hi-C libraries were constructed from fresh leaves of *C. cassia* var. *cassia* according to a previously published procedure [53]. Plant tissues were fixed with formaldehyde and cross-linked DNA digested with DpnII overnight. The sticky ends of the digested fragments were biotinylated and ligated randomly. Chimeric fragments representing the original cross-linked physical interactions were processed into paired-end sequencing libraries. The libraries were subsequently sequenced on Illumina Hiseq X, producing 140.45Gb 2 × 150 base pair paired-end reads.

Total RNA was extracted from each sample by using the RNeasy Plant Mini Kit (Omega, USA). The first-strand cDNAs were synthesized from the total RNA samples using the SuperScript III First-Strand Synthesis System with the oligo(dT) 20 primer (Thermos Fisher Scientific, USA). The RNA-Seq reads were produced on the Illumina NovaSeq 6000 platform.

Genome survey

K-mer method was performed with clean Illumina short data. Jellyfish v2.2.10 [54] was performed to generate a 21-mer frequency distribution, which was then analyzed in GenomeScope 2.0 [55] to estimate the genome size and heterozygosity rate.

De novo genome assembly

To obtain the chromosome-level genome, we integrated the sequencing data obtained with PacBio CCS and Hi-C technology. The raw data of PacBio subreads were filtered to HiFi reads by PBccs v6.4.0 (https://github.com/PacificBiosciences/ccs) and subsequently assembled using Hifiasm v0.16.0 with default parameters [56]. Hi-C reads were subjected to quality control and then mapped to the initial contig assembly of *C. cassia* var. *cassia* using Juicer v1.11.08 [57]. Mis-joins, order and orientation were corrected by the 3D-DNA pipeline v190716 [58]. BUSCO v5.1.2 was used to evaluate the completeness of the chromosome-level genome with the "embryophyte_odb10" database [59].

Genome annotation

Repetitive elements in the genome of *C. cassia*, including LTR, TIR and non-TIR elements, were *de novo* identified, annotated, and classified using Extensive *de novo* TE Annotator (EDTA) v2.0.0 pipeline [60]. Protein-coding genes were initially predicted *ab initio* using the deep learning-based tool Helixer v0.3.1 [61, 62], employing the model 'land_plant_v0.3_a_0080.h5'. To enhance the accuracy and completeness of the gene annotation, RNA-Seq data were integrated into the workflow. RNA-Seq Illumina short-reads derived from different tissues were mapped to the *C. cassia* genome using Hisat2 v2.1.1 [63] and then assembled into transcripts using StringTie v1.3.5 [64]. The resulting transcript assemblies were further validated, and gene structures were predicted using the Program to Assemble Spliced Alignments (PASA) v2.5.1 [65]. Last, gene models identified by Helixer were refined and updated through iterative comparison via PASA. This step ensured improved structural accuracy and annotation completeness by leveraging transcript evidence to correct and extend *ab initio* predictions. For functional annotation, predicted protein-coding genes sequences were aligned to both Kyoto Encyclopedia of Genes and Genomes (KEGG) (www.genome.jp/kegg/) and Gene Ontology (GO) (http://geneontology.org/) via EggNOG-mapper v2.1.10-0 (http://eggnog-mapper.embl.de/).

Comparative genomic and evolutionary analyses

A total of 12 plant species, including six magnoliids (*C. kanehirae*, *Camphora camphora*, *C. burmanni, Persea americana*, *Chimonanthus Salicifolius*, *Liriodendron chinense* and *Aristolochia fimbriata*), three eudicots (*Arabidopsis thaliana*, *Populus trichocarpa*, and *Vitis vinifera*), two monocots (*Zea mays* and *O. sativa*), and one ANA-grade species (*Amborella trichopoda*) were used to infer the phylogenetic tree. Paralogs and orthologs were identified using the OrthoFinder v2.5.2 [66], and protein sequences of 235 single-copy orthologous genes were used to construct a phylogenetic tree. The protein sequences were aligned by MAFFT v7.271 [67], trimmed with trimal v1.4.rev22 [68] and subsequently concatenated by

Phylosuite v1.2.2 [69]. IQ-tree v2.0.3 [70] was used to detect the best-fit amino acid substitution model, based on which, RAxML-NG v.8.2.12 [71] was employed to construct the ML phylogeny with 1,000 bootstrap replicates, and *A. trichopoda* was used as the outgroup. The species tree was then used as an input to estimate divergence time in the MCMCTree program of the PAML package [72]. Multiple fossil times were used for time calibrations from TIMETREE (http://www.timetree.org/). The expansion and contraction of gene families were inferred with CAFE5 [73] based on the chronogram of the above-mentioned 12 plant species.

Whole-genome sequencing and SNP calling

DNAs from 98 leaf tissues were used to construct the 150-bp paired-end libraries in BENAGEN (Wuhan, China), and then were sequenced with the DNBSEQ-T7 platform (MGI, China). The resequencing data was initial quality controled by FastQC v0.11.8 [74], and removed adapters by Trimmomatic v0.38 [75]. After filtering, clean reads were mapped to the reference genome of *C. cassia* using BWA-MEM v0.7.10 [76]. The resulting alignments were sorted to generate bam files using Samtools v1.10 [77], followed by indel realignment, duplicate marking, and base quality recalibration via Picard tools v2.18 (http://picard.sourceforge.net). Calibrated alignments were used for SNP calling with GTX v2.2.1 (https://github.com/aligenomics/gtx-fpga), followed by variant filtration using GATK v4.0.12.0 [78] with these thresholds: QD < 2.0, QUAL < 60.0, MQ < 30.0, FS > 10.0, SOR > 3.0, MQRankSum < -12.5, and ReadPosRankSum < -8.0. Only biallelic SNPs were retained, yielding an initial dataset of 98 individuals and 60,259,942 SNPs. Next, individuals with ambiguous taxonomy or high relatedness were removed, including duplicated or first-degree relatives identified using a kinship coefficient threshold of > 0.354, following standard kinship classifications. PLINK v1.90b6.21 [79] was used to filter out individuals and genetic variants with high missingness using the parameters “--mind 0.2” and “--geno 0.5”. This filtering step resulted in dataset D1, which comprised 88 individuals (include one outgroup *Cinnamomum heyniaum*) and 59,910,409 SNPs. To obtain a more stringently filtered dataset D2, we applied the following additional filters: removal second cousins or closer individuals (kinship coefficient > 0.044, according to standard kinship classifications), exclusion of admixed individuals (ADMIXTURE Q values < 0.7), and filtering of variants with low minor allele frequency using “--maf 0.01”. The resulting dataset D2 contained 43 individuals (include one outgroup *C. heyniaum*) and 867,268 SNPs (Fig. S2).

Population genetics analyses

With the D1 dataset, VCF2Dis v1.46 (https://github.com/BGI-shenzhen/VCF2Dis) was used to calculate the p-distance matrix and build the NJ phylogenetic tree. PCA was performed with PLINK, and Admixture v1.3.0 [80] was used to analyze the population structure, with the number of assumed ancestral populations (K) set from 2 to 4. VCFtools v0.1.16 [81] was used to determine nucleotide diversity (π) for different populations with the following parameters: 50-kb sliding window and 25-kb step size. The same method was also used to calculate the genetic differentiation (F_{ST}) among different groups.

With the D2 dataset, the linkage disequilibrium of “macro” and “cassia” was analyzed by PopLDdecay v3.40 [82]. Balanced subsets were randomly sampled from D2 for sample size-sensitive methods. TreeMix v1.13 [83] was used to test the migration events among different population from 1 to 5. To reduce the computational resource used in the analysis, the D2 subdataset was LD-pruned by PLINK with “--indep-pairwise 50 5 0.2” and vcftools with no missing site before the analysis.

Transcriptome data analyses

RNA-seq raw reads were filtered using fastx-toolkits v0.0.14 (http://hannonlab.cshl.edu/fastx_toolkit/index.html) and then aligned to our assembled *C. cassia* genome with Hisat2, and the quantification of gene expression was calculated with StringTie.

GC-MS Analysis

Frozen biological samples were retrieved from a -80°C freezer and ground in liquid nitrogen to ensure homogeneity. For each sample, approximately 500 mg of material was transferred to a 20 mL headspace vial containing 2 mL saturated NaCl solution and an internal standard. Volatiles were extracted using automated HS-SPME with a 120 μm DVB/CWR/PDMS SPME Arrow (Agilent): samples were incubated at 60°C for 5 min, followed by a 15 min headspace extraction and thermal desorption in the GC inlet at 250°C for 5 min.

GC-MS analysis was conducted on an Agilent 8890 GC coupled to a 7000E mass spectrometer using a DB-5MS column (30 m × 0.25 mm × 0.25 μm) and helium as the carrier gas (1.2 mL/min). The oven program was 40°C (3.5 min), ramping to 100°C at 10°C/min, to 180°C at 7°C/min, and to 280°C at 25°C/min (5 min hold). EI-MS data were collected at 70 eV in SIM mode. Metabolites were identified by matching retention time, fragmentation patterns, and selected ion profiles against a custom-built database, and quantified using target ions normalized to internal standards, following established criteria [84].

LC-MS detection of cinnamaldehyde derivatives

All common chemicals and reagents were obtained from commercial vendors. The standards of *trans*-cinnamaldehyde, 2-hydroxycinnamaldehyde and 2-methoxycinnamaldehyde were purchased from Shanghai yuanye Bio-Technology Co.,Ltd., and the purity of standards exceeded 95%. An UltiMate 3000 HPLC system (Thermo Fisher Scientific,Waltham, MA, USA) fitted with an ACQUITY UPLC BEH C18 Colum (2.1 mm × 100 mm) was coupled to a TSQ Quantum Access masss pectrometer (Thermo Fisher Scientific) operating in positive ion mode. The mobile phase was ultrapure water (A) containing 1% formic acid and acetonitrile (B) with gradient elution. The flow rate was 0.3 ml/min and the injection volume was 2 μL. Method: 0–5 min, 1–2% B; 5–20 min, 2–50% B; 20-20.5 min, 50–90% B; 20.5–23 min,

90% B; 23-23.5 min, 90 – 1% B; 23.5–25 min, 1% B.Additional instrument settings as follows: spray voltage, 3.0 kV; sheath gas (N_2) and auxiliary gas (N_2) flow rates, 30 and 10 (arbi-trary unit), respectively; capillary and auxiliary gas heater temperatures, 320 and 350°C, respectively. The presence of TCA, HCA and MCA in these samples was confirmed via both retention time and MS/MS data, and the concentration was calculated by calibration curves made with standards (Figure S6). Cinnamaldehyde derivatives concentrations in *C. cassia* var. *cassia* and *C. cassia* var. *macrophyllum* across different tissues (bark, leaf, twig) are shown in Table S6.

Cloning of candidate OMT genes

All biosynthetic genes tested in the present study were cloned from the cDNA of *C. cassia* var. *cassia*. CDS for each gene was amplified via PCR using the KOD One PCR master Mix (Toyobo). Oligonucleotide primers were designed to target the 5’ and 3’ homologous regions for subsequent homologous recombination assembly into the appropriate expression plasmid. Amplicon products of PCR were analyzed by gel electrophoresis on a 1% (w/v) agarose gel, and the positive bands were excised and purified using the Omega Gel DNA Recovery Kit (Omega Research). A list of all primers used for CDS cloning in this study is provided in Table S8.

Functional characterization for candidate OMTs in E. coil system

Candidate OMTs were cloned into the pET28a vector, a bacterial expression vector based on maltose-binding protein tag, following the manufacturer’s handbook. The purified plasmids were transformed into Rosetta (DE3) strains of *E. coli* using heat shock transformation. Transformed cells were selected overnight on Lysogeny broth (LB) plates with kanamycin at 37°C. The positive colonies were screened by colony PCR and further cultured in 300 ml of LB containing kanamycin at 200 rpm and 37°C until the optical density at 600 nm (OD600) absorbance reached 0.6 to 0.8, and, then, 0.5 mM isopropyl-β-d-thiogalactopyranoside was added to induce protein expression. After induction, cells were harvested using high-speed centrifugation. The harvested cells were then suspended in 1× PBS buffer and disrupted using sonication. We centrifuged the lysate at 4℃ and collected the supernatant to purify using NI-NTA resin (Qiagen, Hilden, Germany).

Assay for the purified enzyme activity was performed in a 200-µl assay buffer [200 mM NaCl, 1 mM DTT, pH 8.5, 10% glycerol] containing 30 µg of purified proteins, 400 µM S-adenosylmethionine, and 200 µM substrate. The reaction was carried out at 30°C for 3 hours and then extracted with 300 µl of methanol. Each enzyme assay was performed with at least three replicates. LC-MS analysis was performed to detect and quantify the reaction products, as described above.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The raw sequence data reported in this study, including the PacBio HiFi data, Hi-C data and RNA-seq data, have been deposited in the Genome Sequence Archive at the National Genomics Data Center under BioProject number PRJCA056898. Genome assembly and annotation files are available in Figshare (https://doi.org/10.6084/m9.figshare.31152046).

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Key Research and Development Program of China (Grant No. 2023YFA0915800), Guangxi Natural Science Foundation (Grant No. 2025GXNSFAA069567), Guangxi Major Science and Technology Special Project (Grant No. Guike AA24263022), Guangxi Bagui Scholars Special Project (GuangxiTalent2024-01, Kaixiang Li), the National Natural Science Foundation of China (Grant No. 32470245, 32570274, 32400192, and 32300223), and Shenzhen Science and Technology Program (Grant No. JCYJ20241202130723030).

Authors' contributions

L.W. and K.L. conceived and designed the study. L.L., Yu.C, Yi.C and S.S. analyzed the data and prepared figures. Yu.C., X.L. and S.W. performed the wet-lab experiments. Y.Z., C.Z, X.Z and R.J. collected the materials from different regions. L.L, Yu.C, Yi.C, S.S., K.L. and L.W. wrote and reviewed the manuscript. All authors red and approved the final manuscript.

Acknowledgements

Not applicable.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online.

References

1. Pironon S, et al. The global distribution of plants used by humans. Science. 2024;383(6680):293–7.
2. Fornasiero A, Wing RA, Ronald P. Rice domestication Curr Biology. 2022;32(1):R20–4.
3. Cui D et al. Genomic insights on the contribution of introgressions from Xian/Indica to the genetic improvement of Geng/Japonica rice cultivars. Plant Commun, 2022. 3(3).
4. Guo D et al. A pangenome reference of wild and cultivated rice. Nature, 2025: pp. 1–10.
5. Wang W, et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature. 2018;557(7703):43–9.
6. Liu Y, et al. Pan-genome of wild and cultivated soybeans. Cell. 2020;182(1):162–76. e13.
7. Dubcovsky J, Dvorak J. Genome plasticity a key factor in the success of polyploid wheat under domestication. Science. 2007;316(5833):1862–6.
8. Razifard H, et al. Genomic evidence for complex domestication history of the cultivated tomato in Latin America. Mol Biol Evol. 2020;37(4):1118–32.
9. Cao K, et al. Genome-wide association study of 12 agronomic traits in peach. Nat Commun. 2016;7(1):13246.
10. Cornille A, et al. The domestication and evolutionary ecology of apples. Trends Genet. 2014;30(2):57–65.
11. Davis CC, Choisy P. Medicinal plants meet modern biodiversity science. Curr Biol. 2024;34(4):R158–73.
12. Haw SG. Cinnamon, cassia and ancient trade. J Anc History Archaeol, 2017. 4(1).
13. Zheng Z, et al. Genetic diversity analysis of Cinnamomum cassia Presl and Cinnamomum cassia Presl var. macrophyllum Chu based on ISSR. J Holist Integr Pharm. 2024;5(1):19–26.
14. Derks A, Turner S, Hạnh NT. *The taste of cinnamon.* ant Fr, 2022: p. 94.
15. Leela N. *Cinnamon and cassia*, in *Chemistry of spices*. 2008, CABI Wallingford UK. pp. 124–145.
16. Gao H, et al. Transcriptomics and metabolomics analyses reveal the differential accumulation of phenylpropanoids between Cinnamomum cassia Presl and Cinnamomum cassia Presl var. macrophyllum Chu. Ind Crops Prod. 2020;148:112282.
17. Liang X, et al. Comparison of leaf anatomical structure of Cinnamomum cassia and C. cassia var. macrophyll. J Cent South Univ Forestry Technol. 2018;38(11):29–34.
18. Muresan ML, Muresan E. *CHEMICAL COMPOSITION AND ACTIVITY OF THE ESSENTIAL OIL FROM 25 CINNAMOMUM SPECIES. A MINI-REVIEW OF THE LITERATURE.* Studia Universitatis Vasile Goldis Arad. Seria Stiintele Vietii (Life Sci Series). 2024;34(1):26–31.
19. Zhou X, Feng R, Tu H. From Periphery to Nexus: Decoding Borderland Agency in Reshaping Transnational Fragrance Value Chains-Empirical Evidence from Yunnan's Plant-Based Aroma

Industry. Pac Int J. 2025;8(2):12–26.

20. Xing H. *Study on Antibacterial, Insecticidal and Antioxidant activities of the Essential Oil from Cinnamomum cassia Presl. and Cinnamomum loureirii Nees*, in *College of Life Sciences*. South China Agricultural University: Guangzhou, China; 2020.
21. Ju J, de Oliveira MS, Qiao Y. Cinnamon: A Medicinal Plant and A Functional Food Systems. Volume 31. Springer; 2023.
22. Piyasiri K, Wijeratne M. Comparison of the cultivated area and the production trends of Ceylon cinnamon with the main competitors in the worlds' total cinnamon market. Int J Sci Res Publications. 2016;6(1):476–80.
23. Suriyagoda L et al. *Ceylon cinnamon: Much more than just a spice.* Plants, People, Planet, 2021. 3(4): pp. 319–336.
24. Iqbal H, et al. Two promising anti-cancer compounds, 2-hydroxycinnaldehyde and 2-benzoyloxycinnamaldehyde: where do we stand? Comb Chem High Throughput Screen. 2022;25(5):808–18.
25. Banerjee S, Banerjee S. Anticancer potential and molecular mechanisms of cinnamaldehyde and its congeners present in the cinnamon plant. Physiologia. 2023;3(2):173–207.
26. Ye P, et al. Identification of a cinnamoyl-CoA reductase from Cinnamomum cassia involved in trans-cinnamaldehyde biosynthesis. Planta. 2024;259(6):138.
27. Bang HB, et al. Systematic metabolic engineering of Escherichia coli for the enhanced production of cinnamaldehyde. Metab Eng. 2023;76:63–74.
28. Azzahra VO, et al. Synthesis Strategy of Cinnamaldehyde Derivate Compound from Cinnamon Bark Oil (Cinnamomum burmanii) to 2-hydroxycinnamaldehyde. Jurnal Kimia Valensi. 2022;8(1):133–45.
29. Ou S, Jiang N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. 2018;176(2):1410–22.
30. Chaw S-M, et al. Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution. Nat plants. 2019;5(1):63–73.
31. Qin L, et al. Insights into angiosperm evolution, floral development and chemical biosynthesis from the Aristolochia fimbriata genome. Nat Plants. 2021;7(9):1239–53.
32. Dong S et al. The genome of Magnolia biondii Pamp. provides insights into the evolution of Magnoliales and biosynthesis of terpenoids. Hortic Res, 2021. 8.
33. Han P, et al. Exploring genetic diversity and population structure in Cinnamomum cassia (L.) J. Presl germplasm in China through phenotypic, chemical component, and molecular marker analyses. Front Plant Sci. 2024;15:1374648.
34. Gupta N et al. Systematic analysis of antimicrobial activity, phytochemistry, and in silico molecular interaction of selected essential oils and their formulations from different Indian spices against foodborne bacteria. Heliyon, 2023. 9(12).

35. Wahman R, et al. Investigating the potential scabicidal activities of three essential oils: gaining mechanistic insights through in vitro contact bioassay and molecular docking. BMC Complement Med Ther. 2025;25(1):185.
36. Yu P, et al. Effect of peeling and packaging methods on volatile compounds in pepper (Piper nigrum L.) during storage. Int J Food Sci Technol. 2024;59(9):6533–42.
37. Benkhaira N, Ech-Chibani N, Fikri-Benbrahim K. Ethnobotanical survey on the medicinal usage of two common medicinal plants in Taounate Region: Artemisia herba-alba Asso and Ormenis mixta (L.) Dumort. Ethnobotany Res Appl. 2021;22:1–19.
38. Zhu H. A biogeographical study on tropical flora of southern China. Ecol Evol. 2017;7(23):10398–408.
39. Wright IJ, et al. Global climatic drivers of leaf size. Science. 2017;357(6354):917–21.
40. Tang R, Guo Y, Tang Z. Intraspecific variation in leaf morphology of three widespread woody species along climatic gradients. J Plant Ecol. 2024;17(6):rtae091.
41. Upadhyay R, et al. Role of secondary metabolites in plant defense mechanisms: A molecular and biotechnological insights. Phytochem Rev. 2025;24(1):953–83.
42. Lam KC, et al. Structure, function, and evolution of plant O-methyltransferases. Genome. 2007;50(11):1001–13.
43. Rajeswara Rao B. *Genetic diversity, genetic erosion, conservation of genetic resources, and cultivation of medicinal plants*, in *Genetic Diversity and Erosion in Plants: Case Histories*. Springer; 2015. pp. 357–407.
44. Dawson IK, et al. What is the relevance of smallholders' agroforestry systems for conserving tropical tree species and genetic diversity in circa situm, in situ and ex situ settings? A review. Biodivers Conserv. 2013;22(2):301–24.
45. Fang Y, et al. Pan-genome and phylogenomic analyses highlight Hevea species delineation and rubber trait evolution. Nat Commun. 2024;15(1):7232.
46. Tourvas N, et al. Wild and cultivated olive tree genetic diversity in Greece: a diverse resource in danger of erosion. Front Genet. 2023;14:1298565.
47. Sardos J et al. *Painting the diversity of a world's favorite fruit: A next generation catalog of cultivated bananas.* Plants, People, Planet, 2025. 7(1): pp. 263–283.
48. Attikora AJP. *Genetic Diversity and Population Structure of Shea Trees for Sustainable Conservation and Breeding*. in *Plant and Animal Genome Conference (PAG)*. 2025.
49. Rao B. Genetic diversity, genetic erosion, conservation of genetic resources, and cultivation of medicinal plants. Genetic Divers Eros Plants: Case Histories. 2015;8:357.
50. Khoury CK, et al. Crop genetic erosion: understanding and responding to loss of crop diversity. New Phytol. 2022;233(1):84–118.
51. Hou J, et al. The contribution of exotic varieties to maize genetic improvement. Molecular Plant Breeding; 2024. p. 15.

52. Zheng X, et al. A historical review of hybrid rice breeding. J Integr Plant Biol. 2024;66(3):532–45.
53. Xie T, et al. De Novo Plant Genome Assembly Based on Chromatin Interactions: A Case Study of Arabidopsis thaliana. Mol Plant. 2015;8(3):489–92.
54. Marcais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27(6):764–70.
55. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 2020;11(1):1432.
56. Cheng H, et al. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18(2):170–5.
57. Durand NC, et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 2016;3(1):95–8.
58. Dudchenko O, et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356(6333):92–5.
59. Seppey M, Manni M, Zdobnov EM. *BUSCO: assessing genome assembly and annotation completeness.* Gene prediction: methods and protocols, 2019: pp. 227–245.
60. Ou S et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol, 2019. 20(1).
61. Stiehler F, et al. Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning. Bioinformatics. 2020;36(22–23):5291–8.
62. Holst F et al. *Helixer–de novo prediction of primary eukaryotic gene models combining deep learning and a hidden Markov model.* BioRxiv, 2023: p. 2023.02. 06.527280.
63. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60.
64. Pertea M, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290–5.
65. Haas BJ, et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003;31(19):5654–66.
66. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(1):238.
67. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
68. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972–3.
69. Zhang D, et al. PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. Mol Ecol Resour. 2020;20(1):348–55.

70. Minh BQ, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37(5):1530–4.
71. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.
72. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8):1586–91.
73. Mendes FK, et al. CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics. 2020;36(22–23):5516–8.
74. Wingett SW, Andrews S. *FastQ Screen: A tool for multi-genome mapping and quality control.* F1000Research, 2018. 7: p. 1338.
75. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
76. Li H. *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.* arXiv preprint arXiv:1303.3997, 2013.
77. Danecek P, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2):giab008.
78. McKenna A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.
79. Chang CC et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience, 2015. 4(1): p. s13742-015-0047-8.
80. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655–64.
81. Danecek P, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.
82. Zhang C, et al. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics. 2019;35(10):1786–8.
83. Pickrell J, Pritchard J. Inference of population splits and mixtures from genome-wide allele frequency data. Nat Precedings, 2012: pp. 1–1.
84. Yuan H, et al. Development of a widely targeted volatilomics method for profiling volatilomes in plants. Mol Plant. 2022;15(1):189–202.

Figures

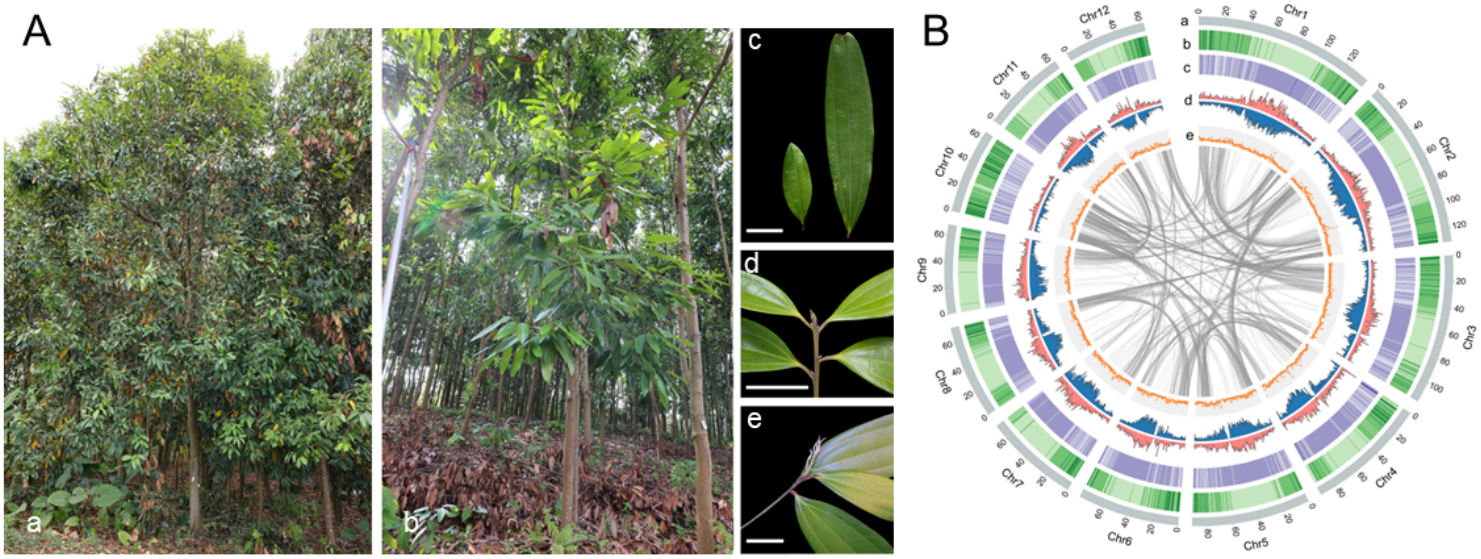


Figure 1

Morphologic and genomic features of *Cinnamomum cassia*.

(A) Images of *C. cassia* and its multiple tissues: (a) Whole plant of Chinese cinnamon (*C. cassia* var. *cassia*), (b) Whole plant of Vietnamese cinnamon (*C. cassia* var. *macrophyllum*), (c) leaves from Chinese (left) and Vietnamese (right) varieties, (d) bud of the Chinese cinnamon; (e) bud of the Vietnamese cinnamon. Scale bar = 5 cm.

(B) Genomic features of *C. cassia* var. *cassia*. Circles from outside to inside: (a) chromosomes, (b) gene density, (c) GC content, (d) LTR/*Copia* density (red) and LTR/*Gypsy* density (blue), and (e) DNA transposable element. Intragenomic syntenies are displayed with connecting lines inside the circos plot.

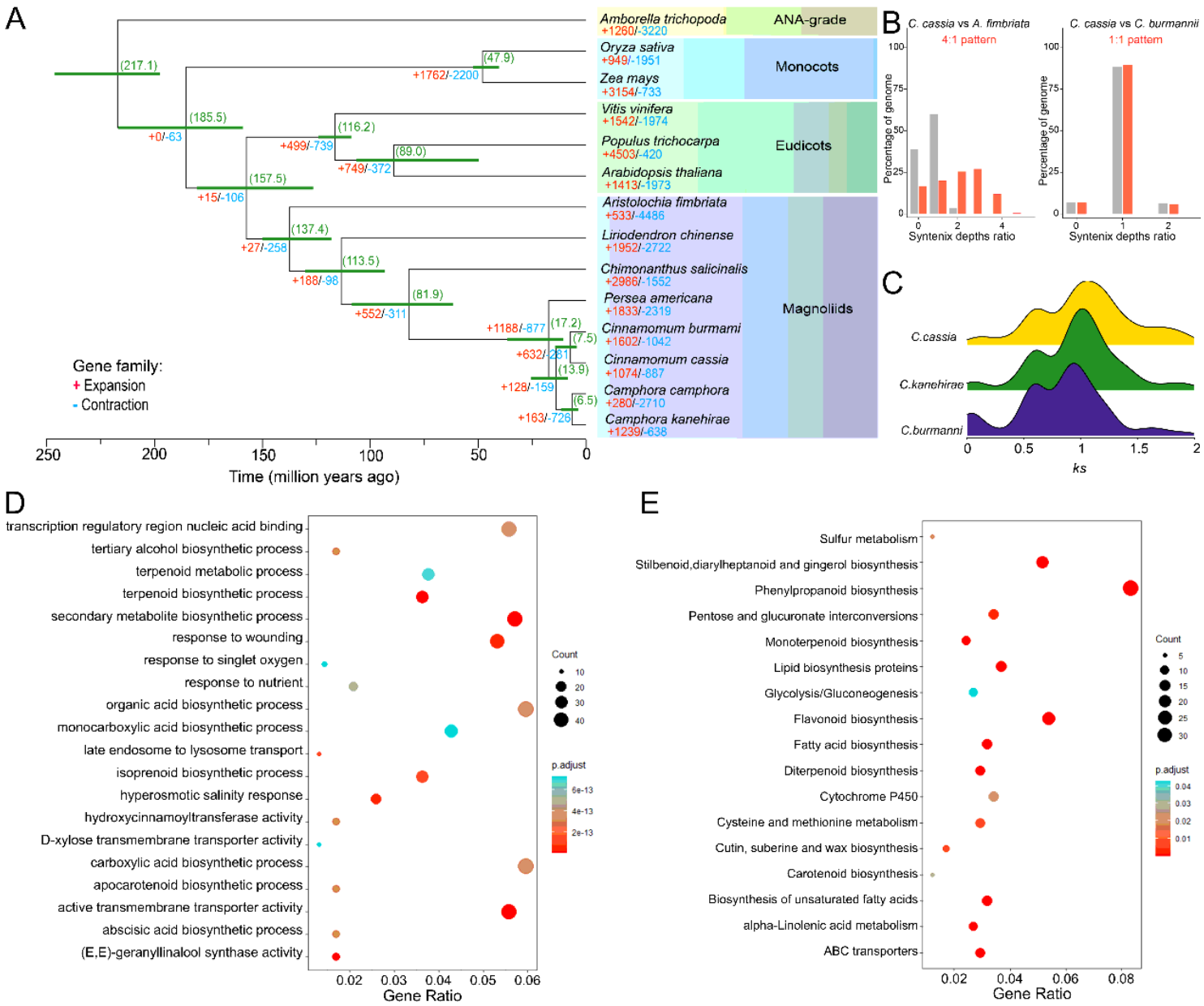


Figure 2

Evolutionary history of *Cinnamomum cassia*.

(A) The phylogenetic tree was constructed based on 334 single-copy orthologous genes of 14 species with one ANA-grade species as the outgroup. Node age and 95% confidence intervals are labeled. +/- signals show the number of gene families that underwent expansion or contraction, respectively.

(B) Syntenic depth histograms for genome-wide collinearity comparisons between *C. cassia* and *A. fimbriata* (left), as well as between *C. cassia* and *C. burmanni*(right). The x-axis represents the syntenic depth ratio (the number of collinear blocks per gene in the other genome), and the y-axis shows the percentage of the genome in each depth class. Grey bars indicate the number of syntenic blocks in the other species per *C. cassia* gene; orange bars represent the reciprocal comparison (the number of *C. cassia* blocks per gene in the other species).

(C) *Ks* distribution of paralogous syntenic gene pairs within each species (*C. cassia*, *C. kanehirae* and *C. burmanni*).

(D) GO enrichment analyses of expanded gene families in *C. cassia*. The top 20 enriched GO terms with adjusted *P*-values < 0.01 are presented. The colors of the bubbles indicate the statistical significance of the enriched GO terms. The sizes of the bubbles represent the number of genes associated with GO terms.

(E) KEGG enrichment analyses of expanded gene families in *C. cassia*. The top 20 enriched KEGG pathways with adjusted *P*-values < 0.01 are presented. The colors of the bubbles represent the statistical significance of enriched KEGG pathways. The sizes of the bubbles indicate the number of genes associated with KEGG pathways.

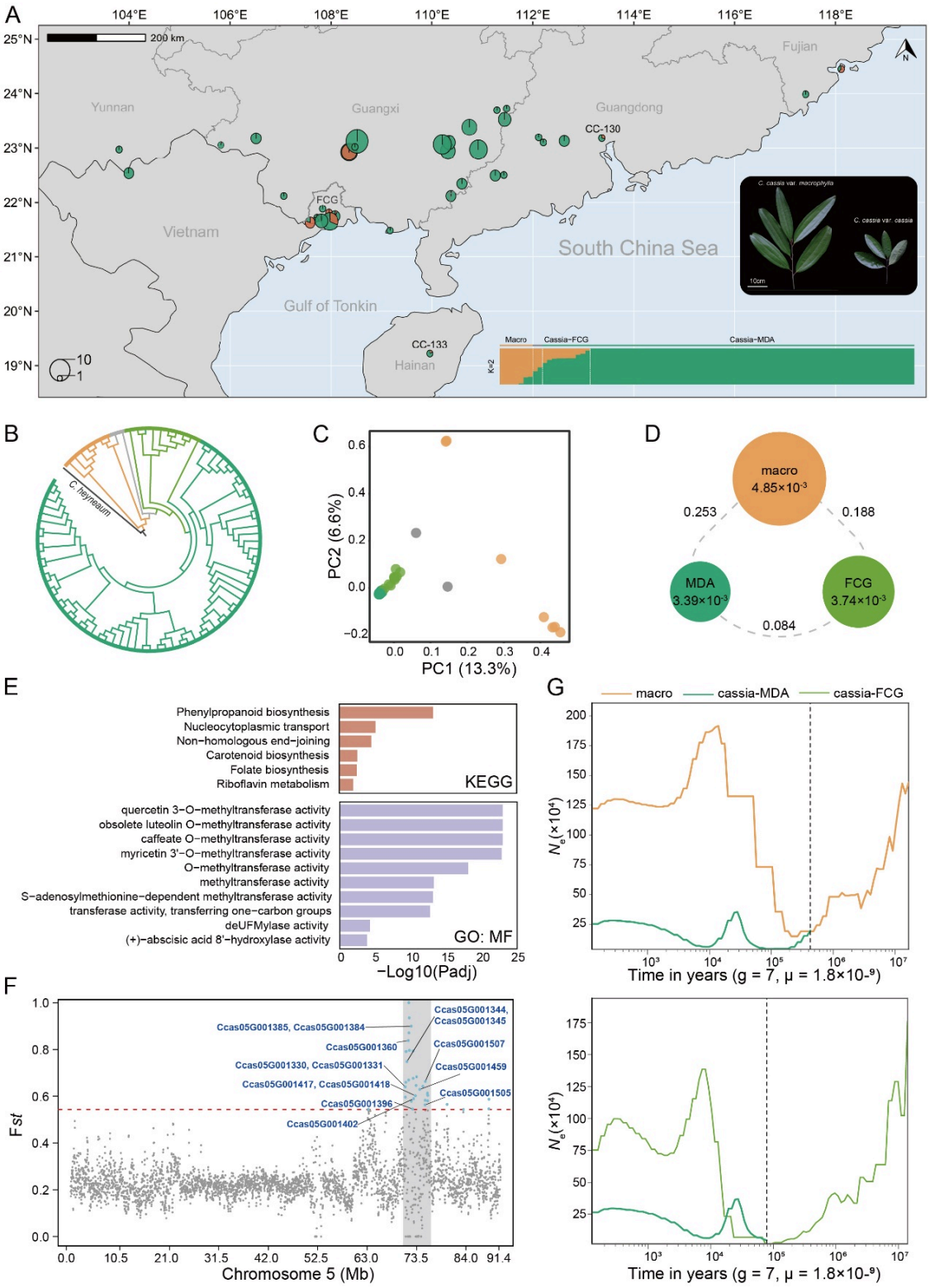


Figure 3

Population genetic analysis of two varieties of *Cinnamomum cassia*. The color code for the populations is consistent in the figure: yellow (macro), light green (cassia-FCG), and dark green (cassia-MDA)

(A) Geographical distribution of the samples (Table S1). Population structure below showed the differentiation of 87 accessions at *K*=2. The position of the pies indicates the sampled locations, while

the radius of the pie denotes the sample size. The different colors in the pies represent ancestral components (according to the revealed population structure at K = 2).

(B) Phylogeny of 87 individuals of *C. cassia* with *C. heyniaum* as the outgroup.

(C) A principal component analysis (PCA) of 87 *C. cassia* individuals. Grey dots represent two admixed individuals excluded from further analyses.

(D) Population differentiation Pairwise fixation index (F_{st}) among populations and the Nucleotide diversity (π) of each population based on 59,910,409 SNPs. The value within each population node corresponds to π, and the value on the connecting line between two populations represents pairwise F_{st}.

(E) Functional enrichment of 381 genes (close to or within the sliding windows with the top 1% F_{st} values) in KEGG pathways and molecular functions (MF) of Gene Ontology. The length of the bars represents the statistical significance of enriched KEGG pathways or GO terms.

(F) Local Manhattan plot of population differentiation on Chromosome 5. Each point represents the average F_{st} calculated for a 25-kb non-overlapping sliding window between the macro and cassia-MDA groups. The red horizontal line indicates the genome-wide top 1% F_{st} threshold (F_{st} = 0.499). Windows exceeding this threshold are highlighted in blue. Gray shaded rectangles denote major differentiation peaks, and OMT genes located within the selected windows or within ±10 kb flanking regions are labeled.

(G) Demographic history of the three *C. cassia* groups via SMC++ analysis. The effective population sizes (N_e) of populations were estimated based on whole-genome resequencing data. The black dashed line represents the estimated divergence time between the two populations.

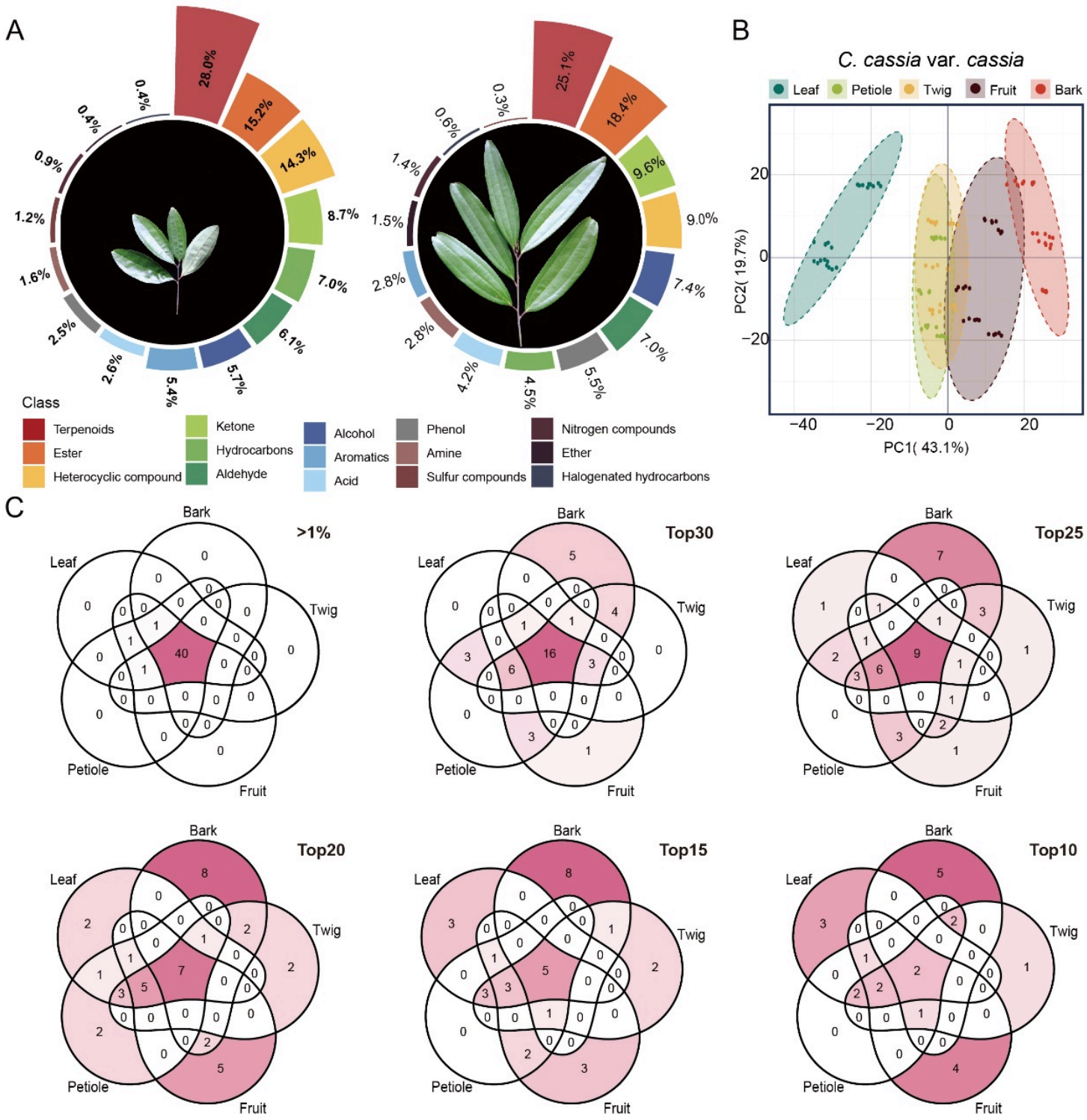


Figure 4

Abundance patterns of volatile metabolites in two *C. cassia* varieties.

(A) Classification and proportion of volatile metabolites detected in *C. cassia* var. *cassia* (left) and *C. cassia* var. *macrophyllum* (right).

(B) Principal component analyses (PCA) of metabolite profiles across different tissues of *C. cassia* var. *cassia*. The ellipses represent 95% confidence intervals for each tissue.

(C) Series of Venn diagrams illustrating the overlap and uniqueness of major volatile metabolites across different tissues of *C. cassia* var. *cassia*: relative content >1%, and the top 30, 25, 20, 15, and 10 volatile metabolites by abundance.

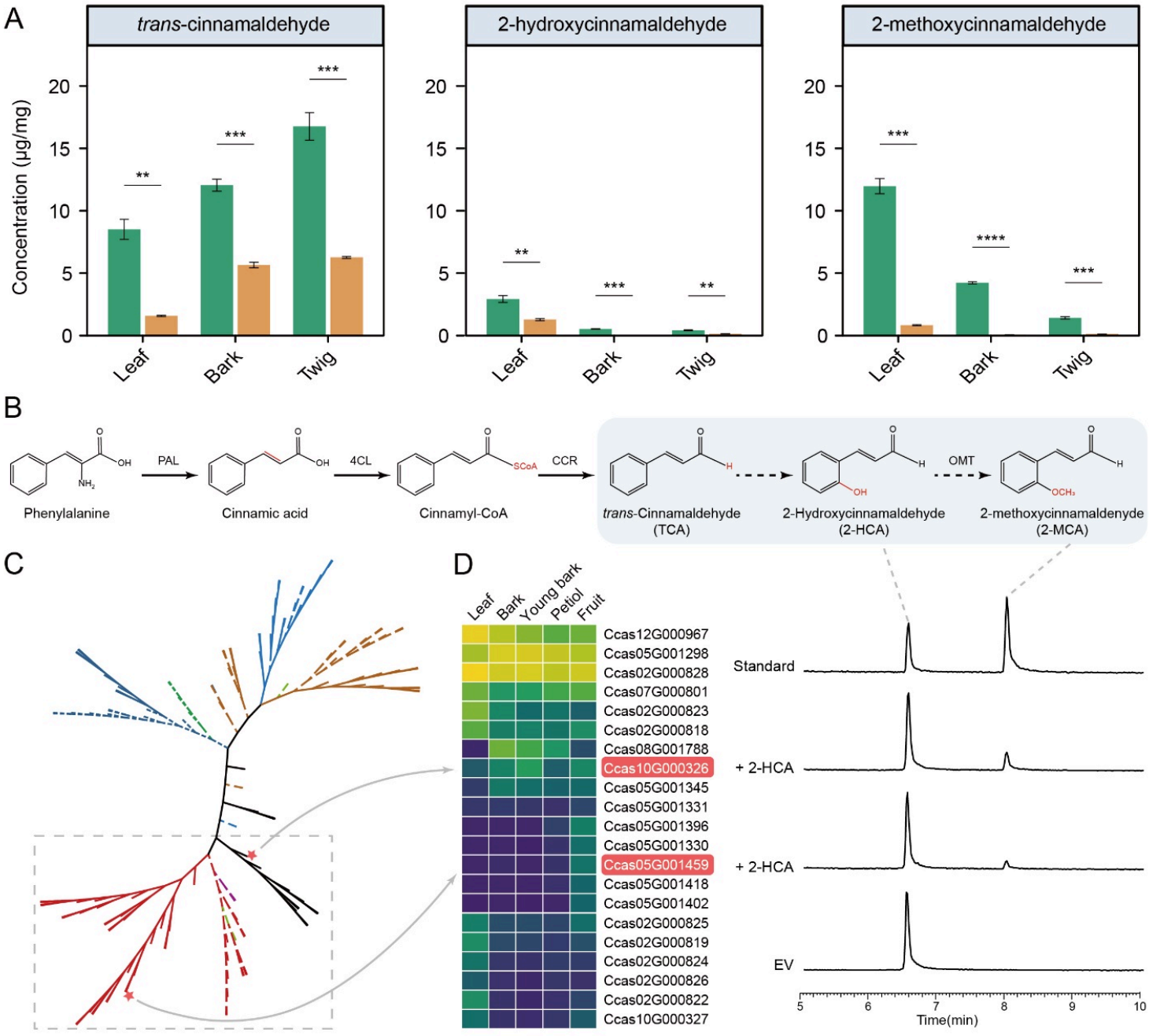


Figure 5

Identification of OMT genes catalyzing the conversion of 2-hydroxycinnamaldehyde to 2-methoxycinnamaldehyde in *C. cassia* var. *cassia*.

(A) Concentrations (μg/mg) of three cinnamaldehyde derivatives (trans-cinnamaldehyde, 2-hydroxycinnamaldehyde, and 2-methoxycinnamaldehyde) across various tissues of Chinese cinnamon **(**green) and Vietnamese cinnamon (yellow)**. Differences between groups were assessed using Welch's two-sample t-test. *P* values were adjusted for multiple comparisons using the Benjamini-Hochberg method. Statistical significance is indicated by asterisks:** $^{**}P < 0.01$; $^{***}P < 0.001$; $^{****}P < 0.0001$.

(B) Proposed biosynthetic pathways for 2-hydroxycinnamaldehyde (2-HCA) and 2-methoxycinnamaldehyde (2-MCA) in *C. cassia*. PAL, Phenylalanine ammonia lyase; 4CL, 4-coumarate:coenzymeA ligase; CCR, Cinnamoyl-CoA reductase; OMT, O-methyltransferase.

(C) Phylogenetic tree of OMTs from *C. cassia* (solid lines) and 52 functionally verified OMTs in plants (dashed lines). Proteins are color-coded by substrate specificity: dark blue, cinnamoyl-CoA OMTs; dark green, carboxylic acid OMTs; yellow, flavonoid/isoflavonoid/flavonol OMTs; blue, benzylisoquinoline/alkaloid OMTs; purpl, other OMTs; red, hydroxycinnamaldehyde/hydroxycinnamic acid/hydroxycinnamyl alcohol OMTs; black, unclassified *C. cassia* OMTs. The dashed box highlights candidate genes for the targeted biosynthetic step.

(D) Expression profiles (TPM) of 21 OMT genes potentially involved in cinnamaldehyde derivative biosynthesis in *C. cassia*. Red boxes indicate candidate genes that catalyze the conversion of 2-HCA to 2-MCA, with corresponding LC-MS chromatograms of enzymatic reactions shown on the right.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- GB2601Supplementaryfigures.docx