Molecular Ecology Resources

WILEY

MOLECULAR ECOLOGY RESOURCES

RESOURCE ARTICLE OPEN ACCESS

A Near Telomere-To-Telomere Genome Assembly of *Coffea arabica* (Mundo Novo) Provides Insights Into Its Secondary Metabolism

Yi Liu[1,2] | Hang Zong[1] | Yaowu Xing[3] | Xi Jiao[2] | Zhuoya Liu[1,2] | Yusheng Niu[1] | Zhiling Yang[3] | Shimeng Liu[2] | Yongqiang Wang[4] | Haodong Zhao[5] | Xianqing Chen[2] | Zhenzhu Li[1,2] | Xiao Wang[2] | Jing Cai[1] | Wen Wang[1] | Zhongkai Wang[1,2]

[1]Shaanxi Key Laboratory of Qinling Ecological Intelligent Monitoring and Protection, School of Ecology and Environment, Northwestern Polytechnical University, Xi'an, Shaanxi, China | [2]Jiaxing Synbiolab Biotechnology co., Ltd., Jiaxing, China | [3]Key Laboratory of Tropical Forest Ecology, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla, China | [4]Sanjie Institute of Forage, Yangling, China | [5]State Key Laboratory of Engineering Biology for Low-Carbon Manufacturing, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin, China

Correspondence: Jing Cai (jingcai@nwpu.edu.cn) | Wen Wang (wenwang@nwpu.edu.cn) | Zhongkai Wang (wangzhongkai@mail.nwpu.edu.cn)

Received: 25 February 2025 | **Revised:** 6 August 2025 | **Accepted:** 18 September 2025

Handling Editor: Joanna Kelley

Funding: This study was supported by National Natural Science Foundation of China, grant no. 32371499, and the New Cornerstone Science Foundation (to W.W.).

Keywords: Arabica coffee | NMTs | Reb M | secondary metabolism | T2T genome assembly | UGTs

ABSTRACT

Arabica coffee (*Coffea arabica*) dominates global coffee production, accounting for over 60% of the world's coffee trade. The Mundo Novo cultivar, predominantly grown in Yunnan, China, represents a significant germplasm resource. However, the absence of a high-quality reference genome has hindered comprehensive genetic research and in-depth investigation of secondary metabolic pathways in Arabica. In this study, we present the first near telomere-to-telomere (T2T) genome assembly of Arabica, achieved through the integration of PacBio HiFi, Oxford Nanopore ultra-long, and Hi-C sequencing technologies, representing the highest-quality Arabica genome to date. Phylogenetic analysis of N-methyltransferases (NMTs), the key enzymes responsible for caffeine biosynthesis, revealed their independent evolution across caffeine-producing clades including coffee, cacao, and tea. Furthermore, GO enrichment analysis of expanded gene families at the Arabica ancestral node, combined with fruit-specific transcriptomic profiling, revealed that glycosyltransferases likely play a critical role in the secondary metabolism of Arabica. Notably, functional characterisation demonstrated that a UGT (uridine diphosphate glycosyltransferase, UGT) from the UGT29 subfamily, which exhibited increased gene copy number in the Arabica subgenome C than its ancestor, can directly convert Rebaudioside A (Reb A) into Rebaudioside M (Reb M) through a single-step enzymatic glycosylation. This direct pathway represents a crucial advancement over conventional multi-UGTs biosynthetic routes of Reb M, which is a highly desirable sweetener whereas with limited natural abundance. Taken together, this study not only provides a valuable genomic resource for studying the unique secondary metabolic processes in *C. arabica* but also accelerates innovative research frontiers for the synthetic biological production of the valuable sweetener Reb M.

Yi Liu, Hang Zong, Yaowu Xing, Xi Jiao, Zhuoya Liu contributed equally to this work.

1 | Introduction

Coffee is a vital economic crop with substantial global trade significance (Tran et al. 2016). Global coffee consumption reaches approximately 165.5 million 60-kg bags annually, with the international coffee trade valued at 3.62 billion US dollars (Charles 2023). The principal varieties employed in coffee production are *Coffea arabica* (commonly called Arabica coffee; hereafter Arabica in brief) and *Coffea canephora* (commonly called Robusta) (Cui et al. 2020; Vidal et al. 2010). *C. arabica* (2n = 4x = 44), an allotetraploid species arising from interspecific hybridization between *C. canephora* and *C. eugenioides* (Lashermes et al. 1999), represents the sole polyploid within the genus *Coffea*. Arabica is renowned for its superior beverage quality compared to Robusta (Jingade et al. 2019) and accounts for approximately 60% of global coffee production (Spinoso-Castillo et al. 2020). The most widely cultivated Arabica variety in China is Mundo Novo (Figueiredo et al. 2019), a cross between Typica and Bourbon (Anthony et al. 2002). Among 130 species of genus *Coffea*, Arabica is uniquely characterised by its distinctive sweetness and pronounced caramel aroma (Nadaleti et al. 2022). Arabica contains approximately 10% polysaccharides and glycosides (Oosterveld et al. 2003), with glycosidic composition determined by the monosaccharide residues and glycosidic bonds, likely involving glycosylation modifications mediated by UDP-glycosyltransferases (UGTs). UGTs could transfer glucose residues from activated uridine diphosphate glucose (UDPG) donors to acceptor molecules (Meech et al. 2019), generating diverse glycoside compounds. In Arabica, several diterpenoid compounds glycosylated by UGTs are believed to contribute to its desirable sensory characteristics (Chu et al. 2016; Hu et al. 2020; Wang et al. 2018). Therefore, exploring the underlying metabolic mechanisms and identifying candidate glycosyltransferases for Arabica flavour enhancement represents an intriguing research avenue.

Caffeine, a significant bioactive compound found in coffee, tea, and cacao, exhibits diverse biological functions including antimicrobial activity (Hewavitharanage et al. 1999; Kim et al. 2011; Nathanson 1984), pest deterrence (Kim and Sano 2008), and suppression of seed germination in competing plants (Pacheco et al. 2008). Caffeine biosynthesis occurs through sequential three-step methylation of xanthine derivatives at positions 7-N, 3-N, and 1-N (Ashihara and Crozier 1999; Ashihara and Suzuki 2004), which is primarily driven by three types of N-methyltransferases (NMTs): xanthine methyltransferase (XMT), 7-methylxanthine methyltransferase (MXMT), and 3,7-dimethylxanthine methyltransferase (DXMT). Phylogenetic evidence suggests that caffeine biosynthetic NMTs have evolved independently at least twice during plant evolution (Denoeud et al. 2014). Additionally, Xia et al. proposed that the NMT genes in tea and cacao may have descended from a common ancestor before their divergence and independent evolution (Xia et al. 2017). However, whether caffeine biosynthetic NMTs from these major caffeinated species share a common origin remains unresolved. Low genome assembly completeness and annotation quality may have hindered comprehensive NMT gene identification across species, potentially leading to biases in evolutionary analyses. An improved Arabica genome assembly may assist in clarifying the evolution of the NMT genes.

Recently, Scalabrin et al. generated a chromosome-level assembly of Arabica 'Bourbon' using ONT long reads and Hi-C data, revealing how chromosomal aberrations and exchanges between homoeologous chromosomes contribute to the genetic diversity in Arabica (Scalabrin et al. 2024). Subsequently, Salojärvi et al. elucidated the diversification history of modern coffee cultivars through chromosome-level assembly of Arabica 'ET-39' and population genomics analysis, establishing a foundation for future genome-assisted breeding of Arabica (Salojärvi et al. 2024). In this study, a new genome assembly for the Arabica 'Mundo Novo' variety was presented, which was expected to fill more gaps compared to the two pioneering Arabica varieties' genomes and achieve a near telomere-to-telomere (T2T) assembly.

We combined the Ultra-long reads from Oxford Nanopore Technology (ONT) and high fidelity reads from PacBio HiFi with the aid of high-throughput chromosome conformation capture (Hi-C). Comparative genomic and phylogenetic analyses were used to detect the genomic evolution of Arabica across related plant lineages. All caffeine biosynthetic NMT genes in Arabica were identified, and the evolution of these genes in coffee, tea, and cacao was clarified. Enrichment analysis of expanded gene families and differential transcriptomic expression were employed to uncover the potential molecular mechanisms underlying the unique flavour of Arabica. Remarkably, a CaUGT from the Arabica was identified, which could directly catalyse the conversion of Reb A into Reb M without the requirement of additional UGTs. This UGT possesses novel glycosylation modification functionality not previously described. Reb M represents a valuable natural sweetener hailed as the world's third sugar source. It has been reported that the synthesis of Reb M from Reb A involves two stages requiring two different UGTs (Li et al. 2024; Liu et al. 2020; Yang et al. 2019). The newly identified UGT in this study with a novel function from Mundo Novo not only sheds light on the genetic basis for the distinctive flavour of Arabica coffee but also unexpectedly presents a promising approach for the industrial production of Reb M. Collectively, this comprehensive genomic and functional analysis advances our fundamental understanding of coffee biology while establishing a robust foundation for both coffee quality enhancement and biotechnological applications.

2 | Materials and Methods

2.1 | Sample Collection and Preparation

Arabica has been widely cultivated in Yunnan, China, for over 150 years. Arabica 'Mundo Novo' was grown in Xishuangbanna Tropical Botanical Garden, Yunnan province, China (101.2768 E, 21.9201 N). Different tissues, including leaves, fruits, stems, and roots, were collected from a single adult plant. Plant samples were then washed with 75% alcohol to remove the surface impurity and congealed in liquid nitrogen. High-quality genomic DNA was isolated from four fresh leaves using the CTAB method. The DNA quality and concentration were assessed using 0.75% agarose gel electrophoresis, NanoDrop One spectrophotometer (Thermo Fisher Scientific), and Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA).

2.2 | Genome Sequencing

For ONT ultra-long sequencing and PacBio HiFi sequencing, two different libraries were constructed and sequenced

to generate enough reads. The libraries with target insert sizes of ~100 kb were built using the Ligation Sequencing Kit (SQKLSK109) and subsequently sequenced using a PromethION sequencer (Oxford Nanopore Technologies, Oxford, UK). The libraries with target insert sizes of ~15 kb were built using NEBNext Ultra DNA Library Prep Kit for Illumina (UK) and subsequently sequenced on a PacBio Sequel II system using Sequencing Primer V2 and the Sequel II Binding Kit 2.1 at Haorui Genomics (Xi'an, China).

For Hi-C sequencing, DNA extracted from the cross-linked samples was digested with the HindIII enzyme and labelled with biotin. The biotin-labelled samples were fragmented to 300–700 bp and subjected to end-repair, adenylation tailing, and universal adapter ligation. The biotinylated fragments were captured using streptavidin beads. Library quality and concentration were assessed using Qubit 2.0 and the Agilent 2100 Bioanalyzer. The effective concentration was validated by qPCR. Paired-end 150 bp (PE150) sequencing was performed using the MGI platform.

For RNA sequencing, an equal quantity of RNAs isolated from each tissue (roots, tender stems, woody stems, young leaves, mature leaves and fruits) was mixed and used to construct libraries using RNA library preparation kits (New England Biolabs, #E7330L). The libraries were subsequently sequenced on the Illumina platform (MGI).

For Next-generation genome sequencing (NGS), the raw data were filtered using Fastp (v0.23.4) (Chen, Zhou, et al. 2018) to remove low-quality reads, adapter contamination, and PCR duplicates. All of the sequencing services were provided by the Haorui Genomics Technology Co. Ltd. (Xi' an, China).

2.3 | Genome Assembly

To evaluate the genome size and heterozygosity of Arabica, k-mer ($k = 21$) analysis was conducted using Meryl (v1.4) (Rhie et al. 2020) and GenomeScope2 (v2.0.1) (Ranallo-Benavidez et al. 2020) with the NGS clean reads. Using ONT Ultra-long reads, HiFi reads, and Hi-C reads, the initial assembly was generated by Hifiasm (v0.19.5-r587) (Cheng et al. 2021) with default parameters (the software utilised in this study employs default parameters unless otherwise specified, which will not be reiterated in subsequent sections). The Hi-C reads were mapped to the initial contig-level assembly using BWA (v0.7.18-r1243) (Li and Durbin 2009). The pairwise interactions between contigs were analysed using 3D-DNA (v190716) (Jia et al. 2022). Juicebox (v2.20.00) was used for manual curation and orientation of contigs (Dudchenko et al. 2017; Durand et al. 2016). A genome-wide Hi-C heatmap was then generated to estimate the quality of the chromosome-scale genome assembly and visualise the interaction matrix of all chromosomes. Gapcloser (v1.2.1) (Xu et al. 2020) was used to fill gaps. Telomeres and centromeres in the Arabica assembly were detected using quarTeT (v1.2.2) (Lin et al. 2023). Given that Arabica is an allopolyploid species, subgenome identification was conducted. Initially, self-alignment of the genome using minimap2 (v2.26-r1175) (Li 2018) was performed to identify homologous chromosomes based on synteny. Then Subphaser (v1.2.6) (Jia et al. 2022) was used to search for the subgenome-specific k-mers and assign homoeologous chromosomes into two subgenomes. Principal component analysis (PCA) was then performed using the prcomp function in R (v4.2.0) to visualise the subgenomic relationships (Katuuramu et al. 2023).

2.4 | Assembly Quality Evaluation

BWA-MEM2 (v2.2.1) (Jung and Han 2022) was used to map the NGS reads to the genome. Minimap2 (v2.28) (Li 2018) was used to map the HiFi reads and Ultra-long reads to the genome. Genome completeness was assessed using BUSCO (v5.0.0) based on the embryophyta_odb10 (version 2020-09-10) database. The completeness and consensus QV of the final genome assembly were assessed using Merqury (v1.3) (Rhie et al. 2020).

2.5 | Genome Annotation

Transposable elements (TEs) in the Arabica genome were annotated by applying the EDTA (v2.0.1) (Ou et al. 2019) pipeline with default parameters. Regions of the genome covered by TEs were soft-masked. For gene annotation, protein sequences from the Asterids clade in UniRef100 (Suzek et al. 2007) were extracted and clustered using MMseqs2 (Release 15-6f452). This protein set contained 825,192 proteins, with complete BUSCO to be 99.90%. Transcriptome sequencing reads were aligned to the reference genome using HISAT2 (v2.2.1) (Kim et al. 2019), and the alignments were converted to BAM format using Samtools (v1.13) (Li et al. 2009). The non-redundant protein set and the BAM-format transcriptome alignment files were used for preliminary gene annotation of the two soft-masked genomes with BRAKER3 (v3.0.3) (Gabriel et al. 2024). The final high-confidence gene set was functionally annotated using eggNOG-mapper (v2.1.12) (Cantalapiedra et al. 2021).

2.6 | Gene Families Analyses

To perform comparative and evolutionary analyses of Arabica, the protein-coding gene sets of 15 species were used to identify orthologous genes, including 12 species of asterids (*Camellia sinensis* (Zhang, Chen, et al. 2021), *Uncaria rhynchophylla* (Hu et al. 2025), *Coffea humblotiana* (Raharimalala et al. 2021), *Scleromitrion diffusum* (Gao et al. 2024), *Gelsemium sempervirens* (Franke et al. 2019), *Marsdenia tenacissima* (Zhou et al. 2023), *Alstonia scholaris* (Chen et al. 2024), *Catharanthus roseus* (Xu et al. 2023), *Gentiana straminea* (Kelsang et al. 2024), *Coffea arabica*, *Coffea canephora* and *Coffea eugenioides* (Salojärvi et al. 2024)), two species of rosids (*Theobroma cacao* (Nousias et al. 2024) and *Arabidopsis thaliana* (Jiao and Schneeberger 2020)), and one basal dicotyledon (*Macadamia integrifolia* (Lin et al. 2022)). Specifically, protein-coding genes of Arabica in this study were split into two subsets according to its subgenome assignment. All protein sequences underwent an all-vs-all Blastp alignment with an e-value of 1e-5. The results were processed using OrthoMCL (v2.0.9) and Mcl (v14-137) to identify orthologous groups. Clustered gene families with gene copies occurring

in at least two species of specified clades, and families with fewer than 100 gene copies, were filtered for CAFÉ software (v4.2.1) to estimate the Lamda value. Gene families with copy numbers exceeding 100 were further analysed using CAFÉ with the previously calculated Lamda value. Families with a *p*-value smaller than 0.05 from both steps were retained, and the results were merged to analyse the expansion and contraction of gene families.

2.7 | Phylogenetic Relationship and Divergence Time

Protein sequences of single-copy orthologs were extracted and performed multiple sequence alignments were performed using MAFFT (v7.471) (Katoh and Standley 2013). PAL2NAL (v14) (Suyama et al. 2006) was utilised to convert protein sequence alignments into codon sequence alignments, optimised by Gblocks (v0.91b) (Castresana 2000), and concatenated into supergenes. The optimal substitution model (LG + I + G) was determined using the ModelFinder program within IQ-TREE (v1.6.12) (Nguyen et al. 2015). Based on this model, a phylogenetic tree was constructed for 15 genomes using IQ-TREE (v1.6.12) (Nguyen et al. 2015) with 1000 bootstrap replicates. MCMCtree (v4.10.7) (Puttick 2019) was used to estimate the divergence times of Arabica relative to other species, using *Macadamia integrifolia* as the outgroup. The divergence times were assessed and calibrated based on two fossil records from the literature (*A. thaliana* vs. *C. roseus* : 111.65–124.3 million years ago; *C. roseus* vs. *C. canephora* : 58.45–79.94 million years ago) (Wang et al. 2021). The resulting phylogenetic tree, annotated with divergence times, was visualised using FigTree (v1.4.4) (Shen et al. 2020).

2.8 | Evolutionary Analysis of the NMT Genes

All available NMT genes were obtained from the UniProt database, including both reviewed and unreviewed entries. Reviewed entries contain manually reviewed sequences with annotation extracted from the literature and curator-evaluated computational analysis, while unreviewed entries contain computationally annotated sequences. All genes were aligned against annotated coding sequences spanning species of tea, cacao, four coffee (Robusta coffee, *C. eugenioides*, *C. humblotiana*, and arabica coffee), and *A. thaliana* . The optimal model (LG + I + G) was determined using ModelFinder. Based on this model, the genes tree and species tree for caffeine biosynthesis were constructed by IQ-TREE (v1.6.12) (Nguyen et al. 2015).

2.9 | Transcriptome Analysis

Transcriptome reads from different tissues, including tender stems, woody stems, young leaves, mature leaves, roots and fruits, were initially quality-filtered using Trimmomatic (v0.39) (Bolger et al. 2014), followed by alignment to the assembled genome using HISAT2 (v2.2.1) (D. Kim et al. 2019). Transcript quantification and expression matrix construction were carried out with StringTie (v2.1.7) (Pertea et al. 2015). Differential gene expression analysis was subsequently performed using DESeq2 (v1.46.0) (Love et al. 2014).

2.10 | GO Enrichment of Specific Gene Set

Gene families of interest and differentially expressed genes in specific tissues were extracted and subjected to GO and KEGG enrichment analyses using the clusterProfiler R package, with functional annotations obtained from Eggnog-mapper software (v2.1.12) (Cantalapiedra et al. 2021). Gene-concept network (CNet) analysis was performed using Enrichplot (v1.26.1) (Yu et al. 2012). For enrichment analysis, background reference terms were filtered using plant-specific GO and KEGG databases. All plant GO terms were downloaded from agriGOv2 (https://systemsbiology.cau.edu.cn/agriGOv2), while plant pathway KO identifiers were retrieved using the R package KEGGREST (v1.46.0) (Tenenbaum and Maintainer 2019).

2.11 | Identification and Functional Validation of UGT Genes in Arabica

Protein-coding genes from 15 species were annotated using the SwissProt database. Genes belonging to the UGT family were identified based on annotation results, and the number of genes corresponding to each UGT subfamily was subsequently quantified across 15 species. The UGT gene was codon-optimised for expression in *E. coli* and synthesised by Wuhan GeneCreate Biological Engineering Co. Ltd. Recombination plasmids were transformed into *E. coli* BL21 (DE3). Recombination strains were precultured overnight at 37°C in 5 mL of LB medium. The seed culture was transferred into 200 mL 2 × YT medium with 100 mg/L kanamycin at 37°C. When the optical density at 600 nm (OD600) reached 0.6–0.8, protein expression was induced by 0.05 mM IPTG (isopropyl-β-D-1-thiogalactopyranoside) at 16°C for 12–16 h. The cells were harvested by centrifugation at 6000 rpm for 10 min and resuspended in 50 mM PBS buffer (pH = 8). The cells were disrupted by ultrasonication on ice for 30 min. The lysate was centrifuged at 8000 rpm for 30 min at 4°C to remove cellular debris. Reb A and UDPG were added to 1 mL crude enzyme and reacted at 37°C for 10 min. The reaction was terminated by adding an equal volume of methanol.

Samples were analysed using a high-precision UPLC system (Waters Technologies, Milford, MA, USA) equipped with an ACQUITY UPLC HSS C18 column (2.1 × 100 mm, 1.8 μm). For HPLC analysis, phase A was acetonitrile and phase B was water with 0.1% formic acid. The following gradient was set as flow rate of 0.6 mL/min: 0–15 min, 25% A linearly increased to 42% A; 15–20 min, 42% A linearly increased to 100% A; 20–25 min, 100% A linearly increased to 5% A and sustained for 10 min; 35–40 min, 25% A. Column temperature was set as 40°C. All samples were detected at a 205 nm wavelength.

Qualitative analysis of each compound was conducted using liquid chromatography-mass spectrometry with a Shimadzu LC-30A and SCIEX Triple TOF 6600 system. Mass spectrometry conditions were as follows: ESI ion source, positive ion IDA detection mode; ion source parameters included an ion voltage of 5500 V, declustering potential of 80 V, source temperature of

550°C, curtain gas at 35 psi, nebulizer gas at 55 psi, and heater gas at 55 psi; the range for the primary scan was m/z 50–1000, and the secondary scan range was m/z 30–950.

2.12 | Elucidation of the Catalytic Mechanism for CaUGT

Enzyme structures were generated using AlphaFold2 (Akdel et al. 2022). The structure exhibiting the highest pLDDT score was selected for subsequent analysis. Substrate structures were retrieved from PubChem (Reb A, CID: 6918840; Reb D, CID: 71773169; Reb I, CID: 92023627; UDPG, CID: 8629). To elucidate possible reasons for the novel catalytic activity of CaUGT, molecular docking was performed using the Watvina (https://github.com/biocheming/watvina). Ligands and protein receptors were prepared using AutodockTools (Morris et al. 2008). The resulting complex structure with the lowest energy was used as the starting conformation for subsequent simulations. Molecular dynamics (MD) simulations were carried out with Gromacs (v2023.2) (Mark James Abraham et al. 2015). Protein calculations were parameterised using the Amber14SB force field (Maier et al. 2015). Ligand molecules were modelled with GAFF (General Amber Force Field) (Sprenger et al. 2015) and Antechamber (Junmei Wang et al. 2001). The systems were enclosed within a periodic boundary octahedral box filled with the TIP3P water model (Price and Brooks 2004). Long-range electrostatic interactions were treated using the Particle Mesh Ewald (PME) method (Abraham and Gready 2011). Each system underwent 5000 steps of the steepest descent and conjugate gradient algorithms for energy minimisation. Production MD simulations were conducted under periodic boundary conditions at 300 K and 1 bar pressure for 100 ns with a 2 fs integration time step. Energy, trajectory, and structural data were recorded every 1 ps. RMSD and RMSF were conducted by Gromacs (v2023.2) (Van Der Spoel et al. 2005).

3 | Results

3.1 | Genome Assembly and Annotation

Based on 105 Gb NGS (Next-Generation DNA sequencing) clean data (Table S1) and k-mer analysis ($k = 21$), the Arabica 'Mundo Novo' genome size was estimated to be 1.16 Gb with 59.52% repeats, and a heterozygosity rate of 4.52% (Figure S1, Table S2). A combination of 49.30 Gb ONT ultra-long reads and 90.25 Gb HiFi reads (Tables S1, S3–S7) was used to construct a preliminary assembly with a genome size of 1.23 Gb (Table S8). Based on 130 Gb Hi-C reads, a total length of 1.19 Gb contigs was further anchored into 22 pseudochromosomes ($2n = 44$), with a mounting rate of 96.75% and a contig N50 of 47.37 Mb (Table 1, Table S9). Principal component analysis (PCA) separated 22 chromosomes into C (SubC) and E (SubE) subgenomes, with PC1 explaining 68.50% of the variance between the two subgenomes (Figure S2A,B). The final genome assembly of Arabica at the chromosome level contains only 14 gaps across 10 chromosomes, while the remaining chromosomes achieved complete T2T assembly without any gaps (Table 1, Figure S3, Tables S10–S12). This assembly exhibits significantly fewer gaps than the two previous Arabica assemblies (Table 1, Table S13). Centromere and telomere locations for all chromosomes were identified, with chromosome sizes ranging from 41.63 to 76.53 Mb (Figure 1A).

The quality of genome assembly was assessed from multiple aspects. First, the genome mapping rates for NGS reads and HiFi reads were 99.41% and 99.65%, respectively, with corresponding mapping coverage of 99.95% and 99.93% (Table S14). Second, BUSCO analysis based on the embryophyta_odb10 database revealed that the assembled genome encompasses 98.70% complete BUSCOs, comprising 120 single-copy and 1473 duplicated BUSCOs (Table S15). Third, Merqury analysis (Rhie et al. 2020) of Illumina sequencing reads yielded a Consensus Quality Value (QV) of 48.95, substantially surpassing the quality benchmark established by the Earth Biogenome Project (EBP) (Lewin et al. 2018). This score demonstrated exceptional assembly accuracy, exceeding the recommended threshold for reference-grade genomes. Additionally, the Hi-C heat map (Figure 1B) displayed strong intra-chromosomal interaction signals with no abnormal inter-chromosomal interactions, supporting the accuracy of the Hi-C-assisted assembly of the Arabica near T2T genome.

A total of 0.84 Gb repeat sequences were identified in the Arabica genome, representing 68.08% of the total genome size. The predominant repeat families were LTR retrotransposons, with Copia elements accounting for 35.88% and Gypsy elements for 8.32% of the genome. DNA transposons constituted 3.95%, while long and short interspersed nuclear elements (LINEs and SINEs) made up a minor proportion of 1.77% (Table S16). The detailed statistics of GC content, gene density, transposable element (TE) density, Copia density, Gypsy density distribution, and chromosomal synteny were displayed in Figure 1C. A total of 71,345 protein-coding genes were annotated using a combination of de novo annotation, homology-based annotation, and transcriptome-based prediction. On average, these genes span 2,331 bp in length, with coding sequence (CDS) averaging 920 bp, intron 4,445 bp, and exon 245 bp. Of these, 69,888 (88.43%) were functionally annotated (Table S17).

3.2 | Comparative Genomic Analysis

The protein-coding genes of 15 species were aligned to identify orthologous genes and assign gene families. 412 single-copy orthologous genes were identified among these species. The phylogenetic tree (Figure 2A) indicated that the Arabica subgenomes C and E are most closely related to *C. canephora* and *C. eugenioides*, respectively, consistent with Arabica's allopolyploid origin. The divergence time estimation showed (Figure 2A) that the *Coffea* species diverged from a common ancestor of the Rubiaceae subfamily approximately 37.00 million years ago (MYA), leading to the emergence of *C. canephora* , *C. eugenioides*, and *C. humblotiana*.

Genetic diversity among different cultivars represents valuable genetic resources for breeding (Li et al. 2023). Our Mundo Novo near T2T genome assembly enables accurate identification of genetic variants among three arabica cultivars with genome sequences available. Taking the Mundo Novo genomes as reference, a total of 157 and 440 structural variations (SVs), comprising 27 and 75 inversions; 5 and 14 translocations; 124 and 339

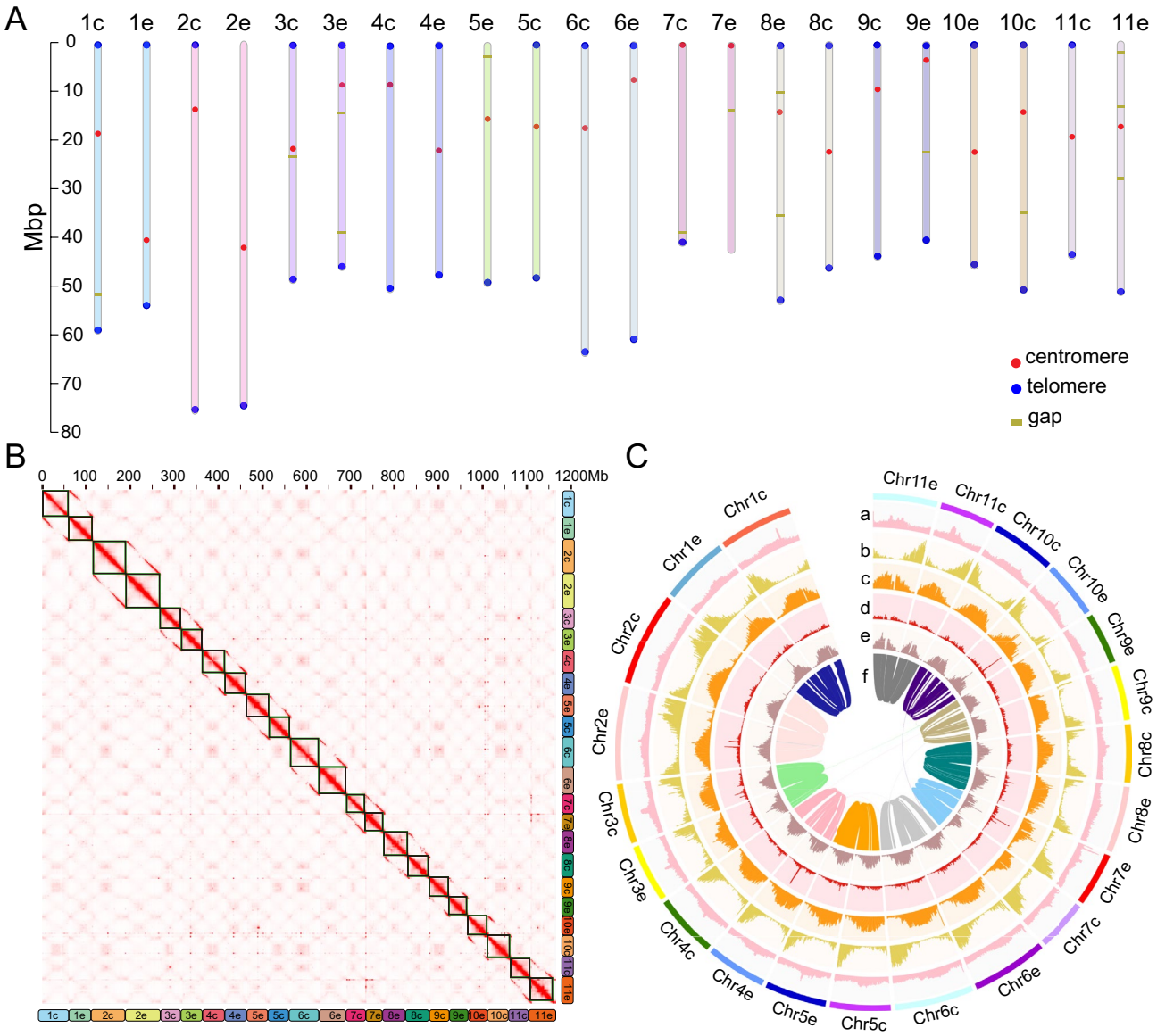

FIGURE 1 | Genome assembly for the Arabica Mundo Novo variety. (A) Distribution of gaps, telomeres, and centromeres on the 22 chromosomes in this Arabica genome. (B) Hi-C interaction heat map between the 22 chromosomes for the Arabica genome. (C) Genome landscape circos plot. The figure illustrates (a) GC content, (b) gene density, (c) transposable element (TE) density, (d) density of Copia, (e) density of Gypsy, and (f) intra-genomic collinearity between the two sets of subgenomes.

TABLE 1 | Comparison of three Arabica varieties' genomes.

Category	Mundo Novo	Bourbon	ET-39
Assembly size (Gb)	1.23	1.30	1.20
Number of contigs	33	175	118
Longest contig (bp)	76,530,765	39,538,567	59,034,463
Contig N50 (Mb)	47.37	9.70	30.00
Number of gaps	14	153	96
Number of telomeres	40	38	22
GC content (%)	37.39	37.00	37.50
QV score	48.95	NA	NA
BUSCO (%)	98.70	98.70	98.60

Abbreviation: NA: not available.

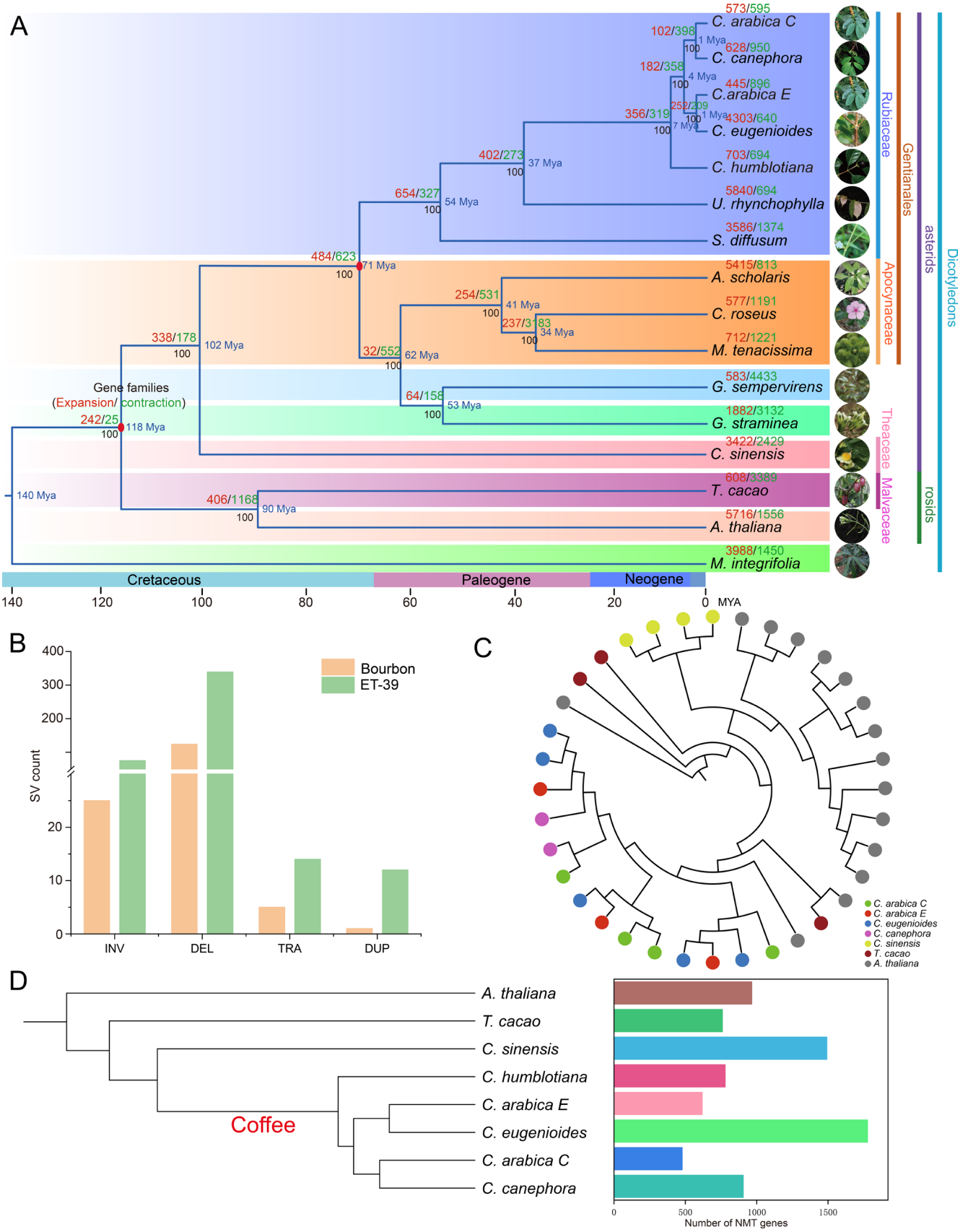

FIGURE 2 | Phylogenetic and evolutionary analysis of the Arabica genome. (A) Phylogenetic tree of 15 species. Gene family expansions and contractions are indicated in red and green colours, respectively. The images of the species used in the figure are sourced from the website https://www.inaturalist.org/. (B) Histogram of SV number and type for three arabica genomes. INV, inversions; DEL, deletions; TRA, translocations; DUP, duplication. (C) Gene tree of all reviewed NMT genes in coffee, tea, cacao and *A. thaliana* . (D) The species tree based on all genes encompassing NMTs among 7 species. The bar chart on the right represents the number of NMTs in each species.

deletions; 1 and 12 duplications were identified in the Bourbon and ET-39 genomes, respectively (Figure 2B, Figure S4A,B, Tables S18, S19).

3.3 | Independent Evolution of Caffeine Biosynthesis-Related NMTs

Caffeine biosynthesis involves three types of N-methyltransferases (NMTs): xanthine methyltransferase (XMT), 7-methylxanthine methyltransferase (MXMT), and 3,7-dimethylxanthine methyltransferase (DXMT) (Ashihara and Suzuki 2004). Previous phylogenetic studies of NMTs involved in caffeine biosynthesis in tea, cacao, and coffee have typically covered limited species (Denoeud et al. 2014; Xia et al. 2017; Zhang, Chen, et al. 2021). Leveraging the chromosome-level genome assemblies of closely related species from recent studies (Nousias et al. 2024; Raharimalala et al. 2021; Salojärvi et al. 2024; Scalabrin et al. 2024; Zhang, Chen, et al. 2021) and our near T2T assembly of Arabica in this study, we were allowed to broadly uncover critical insights into the evolution of NMT genes involved in caffeine biosynthesis from coffee, tea, and cacao. All available NMT genes were retrieved from the UniProt database, including both reviewed and unreviewed ones. All genes were aligned against annotated coding sequences spanning species of tea, cacao, four coffee species (Robusta coffee, *C. eugenioides*, *C. humblotiana*, and Arabica coffee), and *A. thaliana* . As a result, a total of 7 reviewed and 351 unreviewed NMT homologues were identified in Arabica (Tables S20, S21). Phylogenetic analyses showed that the reviewed NMT homologues from tea and cacao form distinct lineages separate from coffee (Figure 2C). The gene tree of all reviewed NMT homologues and the species tree based on all homologues (both reviewed and unreviewed) among 7 species demonstrated that caffeine biosynthetic NMTs from cacao diverged first among these species, forming a sister lineage distinct from tea and coffee. Four *coffea* species, represented by 5 nodes in Figure 2D, clustered together as expected. Importantly, two subgenomes of Arabica clustered distinctly with their ancestral progenitors, respectively. Our results revealed three independent evolutions of caffeine biosynthesis-related NMT homologues among the three caffeinated clades (coffee, tea, and cacao).

3.4 | Glycosyltransferases Play an Important Role in Arabica's Secondary Metabolism

Arabica coffee is highly prized for its exceptional flavour. However, the molecular and genetic mechanisms underlying its unique taste remain largely unexplored. To investigate the potential basis of this distinct flavour, GO enrichment analysis of expanded gene families of Arabica and its ancestors was performed. The results revealed significant enrichment of glycosyltransferases and glycoside biosynthesis, as evidenced by the presence of relevant GO terms and KEGG pathways (Figure 3A, Figure S5, Tables S22, S23). These findings suggested that glycosyltransferases may play a crucial role in the secondary metabolism of coffee. Further functional clustering of GO-enriched terms using CNet analysis showed that a significant portion of the expanded genes possess glucuronosyltransferase activity (Figure 3B). In Arabica subgenome C, both GO and KEGG enrichment analyses of expanded gene families revealed a significant enrichment of genes associated with glycosyltransferase functions (Figures S6, S7, Tables S24, S25), reinforcing the importance of this function in shaping the unique metabolic traits and flavour characteristics of Arabica.

Additionally, transcriptomic data from tender stems, woody stems, young leaves, mature leaves, roots, and fruits of Arabica were further analysed. Three-dimensional PCA demonstrated tight clustering of biological replicates within the same group and clear separation between different tissue types, confirming the sampling reproducibility (Figure S8). Tissue-specific expression patterns revealed that 31.56% (2069/6555) and 58.86% (3858/6555) of differentially expressed genes were uniquely upregulated in fruit and root tissues, respectively (Figure 3C, Table S26). KEGG enrichment analysis of fruit-specific genes indicated significant associations with amino sugar and nucleotide sugar metabolism (ko00520), starch and sucrose metabolism (ko00500), and glycolysis/gluconeogenesis (ko00010) (Figure 3D, Table S27). These metabolic pathways likely contribute to polysaccharide biosynthesis and downstream modifications, with glycosyltransferases potentially mediating critical glycosylation steps. Collectively, the enrichment of glycosyltransferase-related functions in both ancestral expanded gene families and fruit-specific upregulated genes of Arabica suggested a close association between glycosyltransferase activity and the evolution of its distinctive secondary metabolic traits.

3.5 | Characterisation of Glycosyltransferase UGT29 Subfamily in Arabica

In plants, secondary metabolites are reported to be modified through glycosylation by UDP-glycosyltransferases (UGTs) (Wu et al. 2020). Therefore, a comprehensive identification of UGT genes was performed across 15 plant species to investigate their diversity and potential roles in secondary metabolism (Table S28). The UGT gene copy numbers in Arabica, *C. canephora* , and *C. eugenioides* were 209, 135, and 150, respectively. Furthermore, a comparison of the distribution of different UGT families across these species showed a significant enrichment of UGT29 in Arabica, which harbours 25 UGT genes belonging to the UGT29 family. Notably, the subgenome C of Arabica exhibited a remarkably increased UGT29 gene copy number compared to its ancestral species of *C. canephora* . Thus, it is necessary to functionally characterise the UGTs belonging to the UGT29 family in Arabica to elucidate their potential roles. UGTs are involved in the glycosylation of many natural products, with steviol glycosides (SGs) glycosylation in *Stevia rebaudiana* representing one of the most well-known processes. SGs are 250–300 times sweeter than sucrose while containing only 1/300 of the calories of sucrose (Scipioni et al. 2010). Rebaudioside A (Reb A) is one of the main sweetening components in SGs, but it possesses a bitter aftertaste that limits their food industry applications (Gardana et al. 2010; Tian et al. 2022). In contrast, Rebaudioside M (Reb M), which lacks a bitter aftertaste and exhibits a sucrose-like taste, is considered a next-generation natural sweetener. Unfortunately, Reb M comprises only 0.1% of leaf dry weight compared to 4% for Reb A in *S. rebaudiana* leaves (Ceunen and Geuns 2013; Prakash et al. 2014). Previous studies demonstrated that Reb A conversion to Reb M requires two

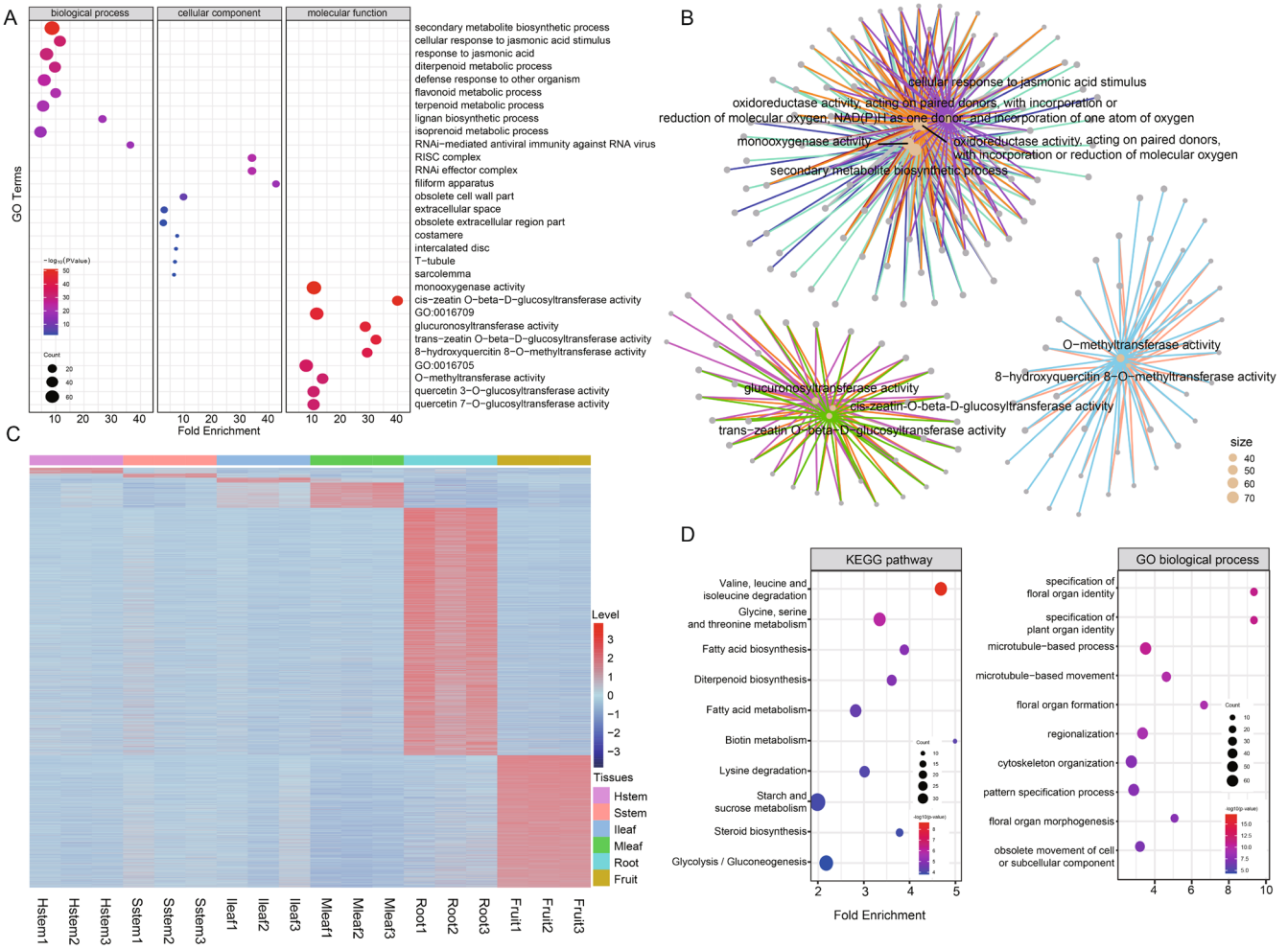

FIGURE 3 | The enrichment analysis for both ancestral expanded gene families and fruit-specific upregulated genes of Arabica. (A) GO (biological process) enrichment analysis of ancestral expanded gene families of Arabica. (B) The cnetplot presenting the network of specific genes from GO terms in ancestral expanded gene families of Arabica. (C) Tissue-specific expression patterns of differentially expressed genes in Arabica. (D) GO (biological process) enrichment analysis and KEGG pathway enrichment analysis of genes in fruit.

sequential steps catalysed by UGT91D2 and UGT76G1, respectively (Yang et al. 2019) (Figure 4A).

To verify UGT function from Arabica, the UGT genes from UGT29 gene families were codon optimised for expression in *E. coli* and inserted into the plasmid pET-28a for heterologous expression. Reb A was catalysed by the recombinant UGT in the presence of UDPG. The retention times for standard Reb A, Reb D, and Reb M were 11.2 min, 9.6 min, and 10.1 min, respectively (Figure 4B). Surprisingly, during functional characterisation, a UGT (gene id: CaraCg5507.t1, hereafter referred to as CaUGT for convenience) from Arabica can directly convert Reb A to Reb M ($[M+H]^+$: *m/z* 1292.548) (Figure 4B, Figure S9). To our knowledge, this is the first report that a single UGT could directly glycosylate Reb A to synthesise Reb M without the synergistic action of additional UGTs. In other words, CaUGT possesses a novel glycosylation capability and significantly shortens the biosynthetic pathway from Reb A to Reb M. The enzyme catalysis conditions were optimised for CaUGT through single-variable experiments. The reaction temperature and pH, which were considered as the key factors influencing the catalytic activity of enzymes, were optimised. The results showed that the reaction at 37°C with pH 8.0 gave the best performance (Figure S10A,B). The influence of 9 different metal ions on CaUGT activity was further explored (Figure S10C); our findings suggest that CaUGT is a metal-dependent enzyme, with Mg^{2+} enhancing activity and Cu^{2+} acting as a potent inhibitor. The yield of Reb M achieved up to 91.65% after catalyzation for 2 h (Figure S10D) under optimal reaction conditions, indicating a promising application potential of the CaUGT in the biosynthesis production of Reb M. These experimental results provided supporting evidence that the UGT29 subfamily likely plays a significant role in the secondary metabolism of Arabica.

Given the CaUGT's unusual glycosylation capability, an analysis of its catalytic mechanism was conducted. We hypothesise that the enzyme initially catalyses the conversion of Reb A to Reb D or Reb I (So far, there is no research indicating that Reb I can be converted into Reb M, and a hypothetical assumption was merely made here), which is then rapidly converted to Reb M. CaUGT was compared with the known glycosyltransferase PaGT3 (PDB ID: 7VEJ) (Maharjan et al. 2022). Our analysis reveals that CaUGT exhibits a typical GT-B fold characterised by two β/α/β Rossmann domains. The N-terminal domain is less

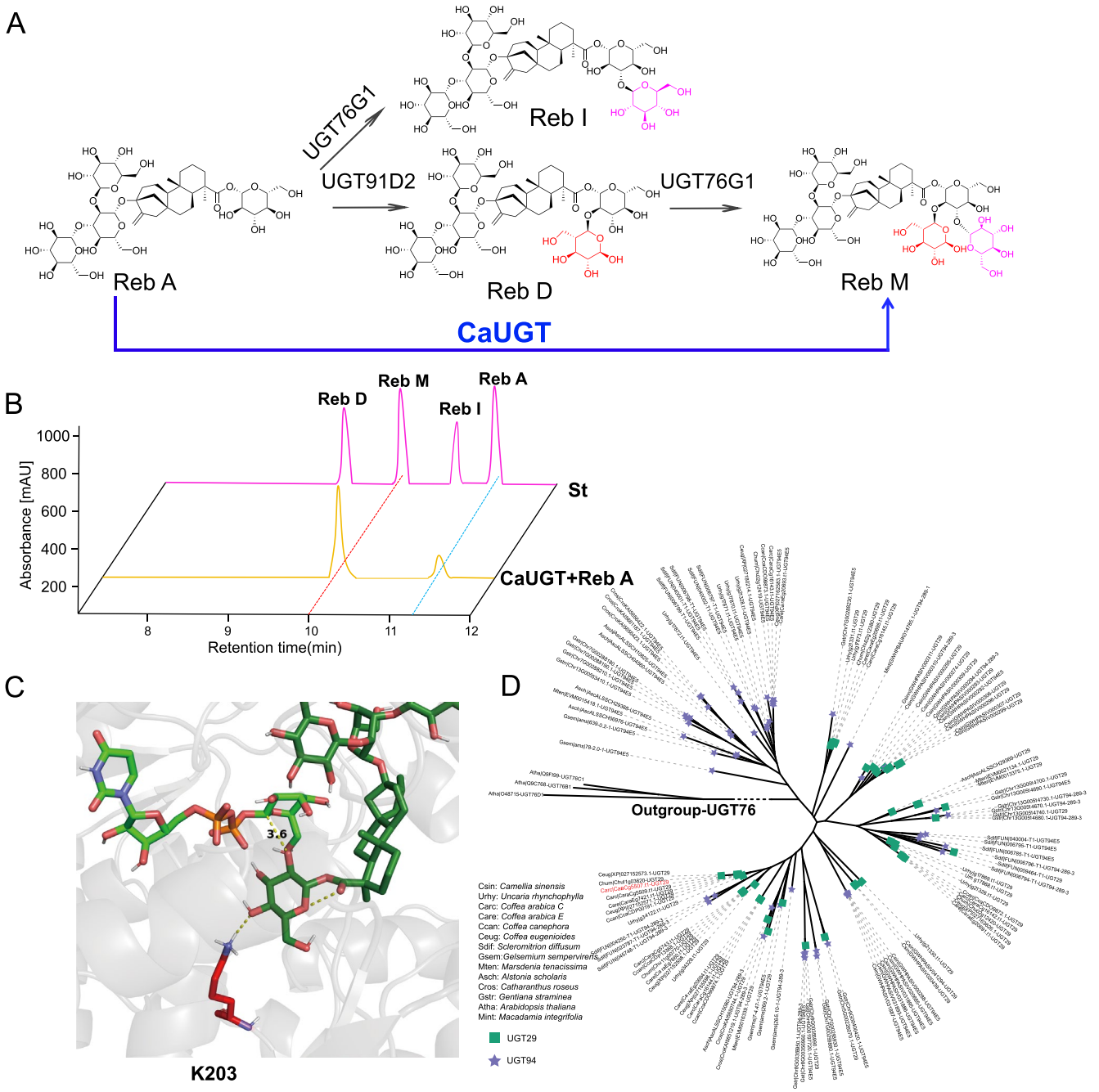

FIGURE 4 | Bioynthesis of Rebaudioside M and phylogenetic analysis of UGT29 and UGT94 across species. (A) Reactions catalysed by UGTs in the Rebaudioside biosynthesis pathway. Rebaudioside was abbreviated as Reb. UGT76G1 and UGT91D2 are known for the biosynthesis of Reb M (Yang et al. 2019; Zhang, Tang, et al. 2021), CaUGT is a novel UGT discovered in this study which can directly catalyse Reb A to Reb M. (B) HPLC-PDA chromatograms (absorbed at 205 nm) of standards (pink) and in vitro reactions catalysed by CaUGT using Reb A as substrate (yellow). St: Standards, Reb D: Rebaudioside D, Reb M: Rebaudioside M, Reb A: Rebaudioside A, Reb I: Rebaudioside I (the isomers of Reb D). (C) Influence of residue K203 on β-1,2 glycosidic bond formation. (D) Phylogenetic analysis of UGT29 and UGT94 across species.

conserved and contains flexible loop structures, which provide a more extensive substrate-binding pocket. The C-terminal domain is more conserved, possessing a 42-amino acid conserved sequence called 'PSPG-box consensus sequence' (PSPG box: plant secondary product glycosyltransferase box) (Hughes and Hughes 1994) (Figure S11A). Based on previous studies (Yang et al. 2019; Zhang, Tang, et al. 2021), the residue H19 was proposed to deprotonate the sugar acceptor to form a nucleophile and attack UDP-glucose to form β-1,2 or β-1,3 glycosidic bonds, while residue D118 stabilises proton transfer. Molecular docking and dynamics simulations revealed that both β-1,2 and β-1,3 glycosidic bond formations are observed during catalysis. A previous study (Liu et al. 2020) suggests that β-1,3 glycosidic bonds in steviol glycosides are typically formed by UGT76G1, while β-1,2 bonds are formed by UGT91D2. Structural comparison (Figure 4C) revealed that CaUGT and UGT91D2 both contain a residue K203 that forms a hydrogen bond with the substrate's oxygen at position four, positioning the sugar ring for

β-1,2 glycosidic bond formation. The hydrophobic interactions between residues L126 and L379 of UGT76G1 and the substrate facilitate the fixation of the substrate on both sides, which is crucial for its proper orientation. Similarly, residues L120 and L362 in CaUGT fulfil this role for β-1,3 glycosidic bond formation (Figure S11B). Consequently, CaUGT may thus be able to facilitate the formation of β-1,2-glycosidic bonds as well as β-1,3-glycosidic bonds. The distance between the UDP-glycosyl donor and the β-1,2 and β-1,3 glycosidic bonds in Reb A was investigated (Figure S11C). The simulation results showed a greater tendency for the formation of the β-1,3-glycosidic bond intermediate Reb I compared to the β-1,2-glycosidic bond product Reb D. This implies that CaUGT initially catalyses the formation of β-1,3 glycosidic bonds, generating intermediate Reb I, which is subsequently followed by the formation of β-1,2 glycosidic bonds to yield Reb M. In summary, CaUGT exhibits characteristic structural features typical of glycosyltransferase and demonstrates the capability to catalyse both β-1,2 and β-1,3 glycosidic bond formations. Inspection of the catalytic structure shows that the enzyme contains key residues for both β-1,2- and β-1,3-glycosidic bond formations. Residue K203 is crucial for β-1,2-glycosidic bond formation by forming hydrogen bonds with the substrate, thereby fixing the sugar ring and enabling UDPG attack. On the other hand, residues L120 and L362 are key for β-1,3-glycosidic bond formation, stabilising the substrate conformation through hydrophobic interactions. These findings provide new insights into the glycosyltransferase mechanism.

4 | Discussion

Coffee is one of the world's most economically important crops, with Arabica contributing approximately 60% of global coffee production. In this study, we employed a combination of sequencing technologies to generate a near telomere-to-telomere (T2T) genome assembly of the Arabica Mundo Novo variety. This high-resolution genome assembly provides a critical resource for elucidating genetic diversity and facilitating comprehensive investigations of secondary metabolic processes in coffee species.

Caffeine, synthesised by three types of NMTs, plays a crucial role in determining coffee quality and contributes to the energising effect that characterises most coffee products. Evolutionary analysis of NMTs offers valuable insights into the molecular evolution of caffeine biosynthesis. Phylogenetic tree across seven species revealed that caffeine biosynthetic NMTs from cacao diverged first, forming a distinct lineage from tea and coffee. The NMTs in coffee underwent the most recent diversification among these species, clarifying the independent evolution of caffeine synthetic NMTs in the coffee, tea, and cacao.

Arabica coffee is known for its unique flavour profile, a primary factor in its market dominance. However, the potential molecular mechanisms underlying its distinctive taste remain largely unexplored. In this study, GO enrichment analysis of expanded gene families across *Coffea* species indicated that glycosyltransferases may play a critical role in the secondary metabolism of coffee species, potentially contributing to the characteristic flavour of Arabica. Consistently, the enrichment of fruit-specific expression further suggested that glycosylation mediated by glycosyltransferases may be a key factor influencing the unique flavour profile of Arabica. The formation of glycosides in plants relies on the glycosylation modifications mediated by UGTs. Our comprehensive analysis of UGTs across coffee species, including robusta coffee, *C. eugenioides*, *C. humblotiana*, and arabica coffee, revealed intriguing patterns. Notably, Arabica subgenome C exhibits a remarkable increase in UGT29 family gene copy number compared to its ancestral species. It was proposed that the UGT29 family may play a crucial role in the flavour of arabica coffee. Remarkably, the examination of these UGTs suggests that a CaUGT from the UGT29 family identified in arabica coffee is capable of catalysing the direct conversion from Reb A to Reb M in a single enzymatic reaction without the synergistic action of any other UGT. This unexpected finding reveals a new pathway for the biosynthesis of Reb M, which has been considered the next generation of sugar substitutes due to its low-calorie, superior sweetness (200–350 times more than sucrose) (Chen, Sun, et al. 2018) and organoleptic properties (López-Carbón et al. 2019). Given the low concentration of Reb M in natural sources, our identified strategy for heterologous biosynthesis represents a significant advancement in potential industrial-scale production for Reb M. Furthermore, the potential roles of CaUGT in the biosynthesis of glycosidic compounds in arabica coffee remain to be explored. The UGT gene clusters in Arabica may serve as valuable targets for molecular breeding, potentially enhancing coffee variety quality. Interestingly, gene family analysis revealed that the UGT29 and UGT94 subfamilies co-localised within the same cluster (cluster267). Phylogenetic analysis of genes within this cluster, using Arabidopsis UGT76 as an outgroup (Figure 4D), demonstrated a close evolutionary relationship between UGT29 and UGT94 subfamilies, with high sequence similarity and homology observed among members of both groups. These findings provide compelling evidence that UGT29 may have originated from the UGT94 subfamily, offering new insights into the molecular basis underlying Arabica's distinctive flavour profile and highlighting the evolutionary significance of UGTs in plant secondary metabolism.

In conclusion, the near T2T genome assembly of Arabica in this study not only provides novel genome resources for coffee improvement and sustainability, but also sheds new insights into the efficient production of an important sweetener, Reb M, through a synthetic biology approach.

Author Contributions

Z.W., Y.L., and W.W. supervised the project; Y.L., Y.X., and Z.Y. collected the plant samples. Y.L., X.C., H.Z., X.W., and Z.L. carried out the experiments; H.Z., Z.L., Y.L., Y.N., Y.W., and S.L. performed bioinformatics analysis; Y.L., H.Z., Y.N., S.L., and Z.W. wrote the manuscript; W.W. and J.C. gave constructive guidance and revised the manuscript.

Acknowledgements

This project was supported by The New Cornerstone Science Foundation (to W.W.) and the National Natural Science Foundation of China (Grant No. 32371499).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive in National Genomics Data Center, China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences (GSA: CRA016801) that is publicly accessible at https://ngdc.cncb.ac.cn/gsa. The genome assembly and the annotation have been deposited into CNGB Sequence Archive (CNSA) of China National GeneBank DataBase (CNGBdb) with accession number CNP0006915.

References

Abraham, M. J., and J. E. Gready. 2011. "Optimization of Parameters for Molecular Dynamics Simulation Using Smooth Particle-Mesh Ewald in GROMACS 4.5." *Journal of Computational Chemistry* 32, no. 9: 2031–2040.

Abraham, M. J., T. Murtola, R. Schulz, et al. 2015. "GROMACS: High Performance Molecular Simulations Through Multi-Level Parallelism From Laptops to Supercomputers." *SoftwareX* 1: 19–25.

Akdel, M., D. E. Pires, E. P. Pardo, et al. 2022. "A Structural Biology Community Assessment of AlphaFold2 Applications." *Nature Structural & Molecular Biology* 29, no. 11: 1056–1067.

Anthony, F., M. Combes, C. Astorga, B. Bertrand, G. Graziosi, and P. Lashermes. 2002. "The Origin of Cultivated Coffee (*Coffea arabica* L.) Revealed by Molecular Markers." *Theoretical and Applied Genetics* 104: 894–900.

Ashihara, H., and A. Crozier. 1999. "Biosynthesis and Metabolism of Caffeine and Related Purine Alkaloids in Plants." In *Advances in Botanical Research*, vol. 30, 117–205. Elsevier.

Ashihara, H., and T. Suzuki. 2004. "Distribution and Biosynthesis of Caffeine in Plants." *Frontiers in Bioscience* 9, no. 2: 1864–1876.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30, no. 15: 2114–2120.

Cantalapiedra, C. P., A. Hernández-Plaza, I. Letunic, P. Bork, and J. Huerta-Cepas. 2021. "eggNOG-Mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale." *Molecular Biology and Evolution* 38, no. 12: 5825–5829.

Castresana, J. 2000. "Selection of Conserved Blocks From Multiple Alignments for Their Use in Phylogenetic Analysis." *Molecular Biology and Evolution* 17, no. 4: 540–552.

Ceunen, S., and J. M. Geuns. 2013. "Steviol Glycosides: Chemical Diversity, Metabolism, and Function." *Journal of Natural Products* 76, no. 6: 1201–1228.

Charles, S. 2023. "Beyond Coffee: Towards a Circular Coffee Economy." In *International Coffee Organization*, edited by S. Charles. International Coffee Organization.

Chen, H., S. K. Sahu, S. Wang, et al. 2024. "Chromosome-Level *Alstonia scholaris* Genome Unveils Evolutionary Insights Into Biosynthesis of Monoterpenoid Indole Alkaloids." *iScience* 27, no. 5: 109599. https://doi.org/10.1016/j.isci.2024.109599.

Chen, L., P. Sun, F. Zhou, et al. 2018. "Synthesis of Rebaudioside D, Using Glycosyltransferase UGTSL2 and In Situ UDP-Glucose Regeneration." *Food Chemistry* 259: 286–291.

Chen, S., Y. Zhou, Y. Chen, and J. Gu. 2018. "Fastp: An Ultra-Fast All-In-One FASTQ Preprocessor." *Bioinformatics* 34, no. 17: i884–i890.

Cheng, H., G. T. Concepcion, X. Feng, H. Zhang, and H. Li. 2021. "Haplotype-Resolved de Novo Assembly Using Phased Assembly Graphs With Hifiasm." *Nature Methods* 18, no. 2: 170–175.

Chu, R., L.-S. Wan, X.-R. Peng, et al. 2016. "Characterization of New Ent-Kaurane Diterpenoids of Yunnan Arabica Coffee Beans." *Natural Products and Bioprospecting* 6: 217–223.

Cui, L., K. Hanika, R. G. Visser, and Y. Bai. 2020. "Improving Pathogen Resistance by Exploiting Plant Susceptibility Genes in Coffee (Coffea spp.)." *Agronomy* 10, no. 12: 1928.

Denoeud, F., L. Carretero-Paulet, A. Dereeper, et al. 2014. "The Coffee Genome Provides Insight Into the Convergent Evolution of Caffeine Biosynthesis." *Science* 345, no. 6201: 1181–1184.

Dudchenko, O., S. S. Batra, A. D. Omer, et al. 2017. "De Novo Assembly of the *Aedes aegypti* Genome Using Hi-C Yields Chromosome-Length Scaffolds." *Science* 356, no. 6333: 92–95.

Durand, N., J. Robinson, M. Shamim, et al. 2016. "Juicebox Provides a Visualization System for Hi-C Contact Maps With Unlimited Zoom." *Cell Systems* 3: 99–101.

Figueiredo, L. P., F. M. Borém, M. R. Almeida, et al. 2019. "Raman Spectroscopy for the Differentiation of Arabic Coffee Genotypes." *Food Chemistry* 288: 262–267.

Franke, J., J. Kim, J. P. Hamilton, et al. 2019. "Gene Discovery in Gelsemium Highlights Conserved Gene Clusters in Monoterpene Indole Alkaloid Biosynthesis." *Chembiochem* 20, no. 1: 83–87. https://doi.org/10.1002/cbic.201800592.

Gabriel, L., T. Brůna, K. J. Hoff, et al. 2024. "BRAKER3: Fully Automated Genome Annotation Using RNA-Seq and Protein Evidence With GeneMark-ETP, AUGUSTUS, and TSEBRA." *Genome Research* 34: 769–777.

Gao, Y., D. Xu, and Z. Hu. 2024. "Telomere-To-Telomere Genome Assembly of *Oldenlandia diffusa*." *DNA Research* 31, no. 3: dsae012. https://doi.org/10.1093/dnares/dsae012.

Gardana, C., M. Scaglianti, and P. Simonetti. 2010. "Evaluation of Steviol and Its Glycosides in *Stevia rebaudiana* Leaves and Commercial Sweetener by Ultra-High-Performance Liquid Chromatography-Mass Spectrometry." *Journal of Chromatography A* 1217, no. 9: 1463–1470.

Hewavitharanage, P., S. Karunaratne, and N. S. Kumar. 1999. "Effect of Caffeine on Shot-Hole Borer Beetle (Xyleborusfornicatus) of Tea (*Camellia sinensis*)." *Phytochemistry* 51, no. 1: 35–41.

Hu, G., X. Peng, Y. Gao, et al. 2020. "Effect of Roasting Degree of Coffee Beans on Sensory Evaluation: Research From the Perspective of Major Chemical Ingredients." *Food Chemistry* 331: 127329.

Hu, T., L. Duan, L. Shangguan, et al. 2025. "Haploid-Phased Chromosomal Telomere-To-Telomere Genome Assembly of Medicinal Plant Uncaria Rhynchophylla Dissects Genetic Controls on the Biosynthesis of Bioactive Alkaloids." *Plant, Cell & Environment* 48, no. 3: 1932–1946. https://doi.org/10.1111/pce.15257.

Hughes, J., and M. A. Hughes. 1994. "Multiple Secondary Plant Product UDP-Glucose Glucosyltransferase Genes Expressed in Cassava (*Manihot esculenta* Crantz) Cotyledons." *DNA Sequence* 5, no. 1: 41–49.

Jia, K. H., Z. X. Wang, L. Wang, et al. 2022. "SubPhaser: A Robust Allopolyploid Subgenome Phasing Method Based on Subgenome-Specific k-Mers." *New Phytologist* 235, no. 2: 801–809.

Jiao, W. B., and K. Schneeberger. 2020. "Chromosome-Level Assemblies of Multiple Arabidopsis Genomes Reveal Hotspots of Rearrangements With Altered Evolutionary Dynamics." *Nature Communications* 11, no. 1: 989. https://doi.org/10.1038/s41467-020-14779-y.

Jingade, P., A. Huded, B. Kosaraju, and M. Mishra. 2019. "Diversity Genotyping of Indian Coffee (*Coffea arabica* L.) Germplasm Accessions by Using SRAP Markers." *Journal of Crop Improvement* 3: 327–345.

Jung, Y., and D. Han. 2022. "BWA-MEME: BWA-MEM Emulated With a Machine Learning Approach." *Bioinformatics* 38, no. 9: 2404–2413. https://doi.org/10.1093/bioinformatics/btac137.

Katoh, K., and D. M. Standley. 2013. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution* 30, no. 4: 772–780.

Katuuramu, D. N., A. Levi, and W. P. Wechter. 2023. "Genome-Wide Association Study of Soluble Solids Content, Flesh Color, and Fruit Shape in Citron Watermelon." *Plant Genome* 16, no. 4: e20391.

Kelsang, G. A., L. Ni, and Z. Zhao. 2024. "Insights From the First Chromosome-Level Genome Assembly of the Alpine Gentian Gentiana Straminea Maxim." *DNA Research* 31, no. 5: dsae022. https://doi.org/10.1093/dnares/dsae022.

Kim, D., J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg. 2019. "Graph-Based Genome Alignment and Genotyping With HISAT2 and HISAT-Genotype." *Nature Biotechnology* 37, no. 8: 907–915.

Kim, Y.-S., S. Lim, H. Yoda, Y.-E. Choi, and H. Sano. 2011. "Simultaneous Activation of Salicylate Production and Fungal Resistance in Transgenic Chrysanthemum Producing Caffeine." *Plant Signaling & Behavior* 6, no. 3: 409–412.

Kim, Y.-S., and H. Sano. 2008. "Pathogen Resistance of Transgenic Tobacco Plants Producing Caffeine." *Phytochemistry* 69, no. 4: 882–888.

Lashermes, P., M.-C. Combes, J. Robert, et al. 1999. "Molecular Characterisation and Origin of the *Coffea arabica* L. Genome." *Molecular and General Genetics MGG* 261, no. 2: 259–266.

Lewin, H. A., G. E. Robinson, W. J. Kress, et al. 2018. "Earth BioGenome Project: Sequencing Life for the Future of Life." *Proceedings of the National Academy of Sciences of the United States of America* 115, no. 17: 4325–4333.

Li, H. 2018. "Minimap2: Pairwise Alignment for Nucleotide Sequences." *Bioinformatics* 34, no. 18: 3094–3100.

Li, H., and R. Durbin. 2009. "Fast and Accurate Short Read Alignment With Burrows–Wheeler Transform." *Bioinformatics* 25, no. 14: 1754–1760.

Li, H., B. Handsaker, A. Wysoker, et al. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25, no. 16: 2078–2079.

Li, N., Q. He, J. Wang, et al. 2023. "Super-Pangenome Analyses Highlight Genomic Diversity and Structural Variation Across Wild and Cultivated Tomato Species." *Nature Genetics* 55, no. 5: 852–860.

Li, S., S. Luo, X. Zhao, et al. 2024. "Efficient Conversion of Stevioside to Rebaudioside M in *Saccharomyces cerevisiae* by a Engineering Hydrolase System and Prolonging the Growth Cycle." *Journal of Agricultural and Food Chemistry* 72, no. 14: 8140–8148.

Lin, J., W. Zhang, X. Zhang, et al. 2022. "Signatures of Selection in Recently Domesticated Macadamia." *Nature Communications* 13, no. 1: 242. https://doi.org/10.1038/s41467-021-27937-7.

Lin, Y., C. Ye, X. Li, et al. 2023. "quarTeT: A Telomere-To-Telomere Toolkit for Gap-Free Genome Assembly and Centromeric Repeat Identification." *Horticulture Research* 10, no. 8: uhad127.

Liu, Z., J. Li, Y. Sun, P. Zhang, and Y. Wang. 2020. "Structural Insights Into the Catalytic Mechanism of a Plant Diterpene Glycosyltransferase SrUGT76G1." *Plant Communications* 1, no. 1: 100004.

López-Carbón, V., A. Sayago, R. González-Domínguez, and Á. Fernández-Recamales. 2019. "Simple and Efficient Green Extraction of Steviol Glycosides From *Stevia Rebaudiana* Leaves." *Food* 8, no. 9: 402.

Love, M. I., W. Huber, S. Anders, and J. G. b. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data With DESeq2." *Genome Biology* 15, no. 12: 550.

Maharjan, R., Y. Fukuda, T. Nakayama, H. Hamada, S.-I. Ozaki, and T. Inoue. 2022. "Structural Basis for Substrate Recognition in the *Phytolacca americana* Glycosyltransferase PaGT3." *Acta Crystallographica Section D: Structural Biology* 78, no. 3: 379–389.

Maier, J. A., C. Martinez, K. Kasavajhala, et al. 2015. "ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters From ff99SB." *Journal of Chemical Theory and Computation* 11, no. 8: 3696–3713.

Meech, R., D. G. Hu, R. A. McKinnon, et al. 2019. "The UDP-Glycosyltransferase (UGT) Superfamily: New Members, New Functions, and Novel Paradigms." *Physiological Reviews* 99, no. 2: 1153–1222.

Morris, G. M., R. Huey, and A. J. Olson. 2008. "Using Autodock for Ligand-Receptor Docking." *Current Protocols in Bioinformatics* 24, no. 1: 11–40.

Nadaleti, D. H. S., J. C. de Rezende Abrahão, M. R. Malta, C. S. Dos Santos, A. A. Pereira, and G. R. Carvalho. 2022. "Influence of Postharvest Processing on the Quality and Sensory Profile of Groups of Arabica Coffee Genotypesc." *Journal of the Science of Food and Agriculture* 102, no. 15: 6899–6906.

Nathanson, J. A. 1984. "Caffeine and Related Methylxanthines: Possible Naturally Occurring Pesticides." *Science* 226, no. 4671: 184–187.

Nguyen, L.-T., H. A. Schmidt, A. Von Haeseler, and B. Q. Minh. 2015. "IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies." *Molecular Biology and Evolution* 32, no. 1: 268–274.

Nousias, O., J. Zheng, T. Li, et al. 2024. "Three de Novo Assembled Wild Cacao Genomes From the Upper Amazon." *Scientific Data* 11, no. 1: 369. https://doi.org/10.1038/s41597-024-03215-1.

Oosterveld, A., J. Harmsen, A. Voragen, and H. Schols. 2003. "Extraction and Characterization of Polysaccharides From Green and Roasted *Coffea arabica* Beans." *Carbohydrate Polymers* 52, no. 3: 285–296.

Ou, S., W. Su, Y. Liao, et al. 2019. "Benchmarking Transposable Element Annotation Methods for Creation of a Streamlined, Comprehensive Pipeline." *Genome Biology* 20: 1–18.

Pacheco, A., J. Pohlan, and M. Schulz. 2008. "Allelopathic Effects of Aromatic Species Intercropped With Coffee: Investigation of Their Growth Stimulation Capacity and Potential of Caffeine Uptake in Puebla, Mexico." *Allelopathy Journal* 21, no. 1: 39.

Pertea, M., G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, and S. L. Salzberg. 2015. "StringTie Enables Improved Reconstruction of a Transcriptome From RNA-Seq Reads." *Nature Biotechnology* 33, no. 3: 290–295.

Prakash, I., A. Markosyan, and C. Bunders. 2014. "Development of Next Generation Stevia Sweetener: Rebaudioside M." *Food* 3, no. 1: 162–175.

Price, D. J., and C. L. Brooks. 2004. "A Modified TIP3P Water Potential for Simulation With Ewald Summation." *Journal of Chemical Physics* 121, no. 20: 10096–10103.

Puttick, M. N. 2019. "MCMCtreeR: Functions to Prepare MCMCtree Analyses and Visualize Posterior Ages on Trees." *Bioinformatics* 35, no. 24: 5321–5322.

Raharimalala, N., S. Rombauts, A. McCarthy, et al. 2021. "The Absence of the Caffeine Synthase Gene Is Involved in the Naturally Decaffeinated Status of Coffea Humblotiana, a Wild Species From Comoro Archipelago." *Scientific Reports* 11, no. 1: 8119. https://doi.org/10.1038/s41598-021-87419-0.

Ranallo-Benavidez, T. R., K. S. Jaron, and M. C. Schatz. 2020. "GenomeScope 2.0 and Smudgeplot for Reference-Free Profiling of Polyploid Genomes." *Nature Communications* 11, no. 1: 1432.

Rhie, A., B. P. Walenz, S. Koren, and A. M. Phillippy. 2020. "Merqury: Reference-Free Quality, Completeness, and Phasing Assessment for Genome Assemblies." *Genome Biology* 21: 1–27.

Salojärvi, J., A. Rambani, Z. Yu, et al. 2024. "The Genome and Population Genomics of Allopolyploid *Coffea arabica* Reveal the Diversification History of Modern Coffee Cultivars." *Nature Genetics* 56, no. 4: 721–731.

Scalabrin, S., G. Magris, M. Liva, et al. 2024. "A Chromosome-Scale Assembly Reveals Chromosomal Aberrations and Exchanges Generating Genetic Diversity in *Coffea arabica* Germplasm." *Nature Communications* 15, no. 1: 463.

Scipioni, G. P., D. J. Ferreyra, M. G. Acuña, and M. E. Schmalko. 2010. "Rebaudioside A Release From Matrices Used in a Yerba Maté Infusion." *Journal of Food Engineering* 100, no. 4: 627–633.

Shen, Y., N. Yang, Z. Liu, Q. Chen, and Y. J. G. Li. 2020. "Phylogenetic Perspective on the Relationships and Evolutionary History of the Acipenseriformes." *Genomics* 112, no. 5: 3511–3517.

Spinoso-Castillo, J., E. Escamilla-Prado, V. Aguilar-Rincón, et al. 2020. "Genetic Diversity of Coffee (Coffea spp.) in Mexico Evaluated by Using DArTseq and SNP Markers." *Genetic Resources and Crop Evolution* 67: 1795–1806.

Sprenger, K., V. W. Jaeger, and J. Pfaendtner. 2015. "The General AMBER Force Field (GAFF) can Accurately Predict Thermodynamic and Transport Properties of Many Ionic Liquids." *Journal of Physical Chemistry B* 119, no. 18: 5882–5895.

Suyama, M., D. Torrents, and P. Bork. 2006. "PAL2NAL: Robust Conversion of Protein Sequence Alignments Into the Corresponding Codon Alignments." *Nucleic Acids Research* 34, no. 2: W609–W612.

Suzek, B. E., H. Huang, P. McGarvey, R. Mazumder, and C. H. Wu. 2007. "UniRef: Comprehensive and Non-Redundant UniProt Reference Clusters." *Bioinformatics* 23, no. 10: 1282–1288. https://doi.org/10.1093/bioinformatics/btm098.

Tenenbaum, D., and B. Maintainer. 2019. "KEGGREST: Client-Side REST Access to the Kyoto Encyclopedia of Genes and Genomes (KEGG)."

Tian, X., F. Zhong, and Y. Xia. 2022. "Dynamic Characteristics of Sweetness and Bitterness and Their Correlation With Chemical Structures for Six Steviol Glycosides." *Food Research International* 151: 110848.

Tran, H. T., L. S. Lee, A. Furtado, H. Smyth, and R. J. Henry. 2016. "Advances in Genomics for the Improvement of Quality in Coffee." *Journal of the Science of Food and Agriculture* 96, no. 10: 3300–3312.

Van Der Spoel, D., E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. J. Berendsen. 2005. "GROMACS: Fast, Flexible, and Free." *Journal of Computational Chemistry* 26, no. 16: 1701–1718.

Vidal, R. O., J. M. C. Mondego, D. Pot, et al. 2010. "A High-Throughput Data Mining of Single Nucleotide Polymorphisms in Coffea Species Expressed Sequence Tags Suggests Differential Homeologous Gene Expression in the Allotetraploid *Coffea arabica*." *Plant Physiology* 154, no. 3: 1053–1066.

Wang, J., W. Wang, P. A. Kollman, and D. A. Case. 2001. "Antechamber: An Accessory Software Package for Molecular Mechanical Calculations." *Journal of the American Chemical Society* 222, no. 1: 2001.

Wang, J., S. Xu, Y. Mei, et al. 2021. "A High-Quality Genome Assembly of Morinda Officinalis, a Famous Native Southern Herb in the Lingnan Region of Southern China." *Horticulture Research* 8, no. 1: 135. https://doi.org/10.1038/s41438-021-00551-w.

Wang, X., Q. Meng, X. Peng, G. Hu, and M. Qiu. 2018. "Identification of New Diterpene Esters From Green Arabica Coffee Beans, and Their Platelet Aggregation Accelerating Activities." *Food Chemistry* 263: 251–257.

Wu, B., X. Liu, K. Xu, and B. J. Zhang. 2020. "Genome-Wide Characterization, Evolution and Expression Profiling of UDP-Glycosyltransferase Family in Pomelo (*Citrus grandis*)." *Fruit* 20, no. 1: 459.

Xia, E.-H., H.-B. Zhang, J. Sheng, et al. 2017. "The Tea Tree Genome Provides Insights Into Tea Flavor and Independent Evolution of Caffeine Biosynthesis." *Molecular Plant* 10, no. 6: 866–877.

Xu, M., L. Guo, S. Gu, et al. 2020. "TGS-GapCloser: A Fast and Accurate Gap Closer for Large Genomes With Low Coverage of Error-Prone Long Reads." *GigaScience* 9, no. 9: giaa094.

Xu, Z., G. Wang, Q. Wang, et al. 2023. "A Near-Complete Genome Assembly of Catharanthus Roseus and Insights Into Its Vinblastine Biosynthesis and High Susceptibility to the Huanglongbing Pathogen." *Plant Communications* 4, no. 6: 100661. https://doi.org/10.1016/j.xplc.2023.100661.

Yang, T., J. Zhang, D. Ke, et al. 2019. "Hydrophobic Recognition Allows the Glycosyltransferase UGT76G1 to Catalyze Its Substrate in Two Orientations." *Nature Communications* 10, no. 1: 3214.

Yu, G., L.-G. Wang, Y. Han, and Q.-Y. He. 2012. "clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters." *OMICS: A Journal of Integrative Biology* 16, no. 5: 284–287.

Zhang, J., M. Tang, Y. Chen, et al. 2021. "Catalytic Flexibility of Rice Glycosyltransferase OsUGT91C1 for the Production of Palatable Steviol Glycosides." *Nature Communications* 12, no. 1: 7030.

Zhang, X., S. Chen, L. Shi, et al. 2021. "Haplotype-Resolved Genome Assembly Provides Insights Into Evolutionary History of the Tea Plant *Camellia sinensis*." *Nature Genetics* 53, no. 8: 1250–1259.

Zhou, Y., W. Fan, H. Zhang, et al. 2023. "Marsdenia Tenacissima Genome Reveals Calcium Adaptation and Tenacissoside Biosynthesis." *Plant Journal* 113, no. 6: 1146–1159. https://doi.org/10.1111/tpj.16081.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Table S1:** Sequencing data used for *C. arabica* genome assembly and annotation. **Table S2:** The data statistics of the 21-mer analysis and heterozygosity in *C. arabica* geno. **Table S9:** Metrics of the assembly at chromosome level. **Table S12:** Metrics of the *C. arabica* genome. **Table S14:** Mapping statistics of NGS/Ultra Long/HiFi reads to the final. **Table S15:** BUSCO assessment of the *C. arabica* genome assembly. **Table S16:** Statistics of the repeat sequences in the *C. arabica* genome. **Table S17:** Statistics of function annotation of the *C. arabica* . **Table S26:** Statistics of differentially expressed genes in Arabica. **Data S1:** men70053-sup-0002-Tables.xlsx. **Figure S1:** Distribution of K-mer frequencies. The plot is generated using GenomeScope2.0 ($k = 21$, ploidy = 4). **Figure S2:** A. Principal component analysis (PCA) of differential 15-mers. **Figure S3:** Location of Closed or Unclosed Gaps in the *Coffea arabica* . **Figure S4:** Structural variations (SVs) among 3 different arabica. **Figure S5:** KEGG pathway analysis of expanded gene families in Arabica's ancestors. **Figure S6:** GO (biological process) enrichment analysis of expanded gene families in Arabica subgenome C. **Figure S7:** KEGG pathway analysis of expanded gene families in Arabica subgenome C. **Figure S8:** Three-dimensional PCA plot of transcriptomic data from different tissue types. **Figure S9:** LC/MS analysis of product Rebaudioside M generated by CaUGT. Reb M: [M + H]+, 1292.548. **Figure S10:** Optimization of the reaction conditions for CaUGT. A. pH optimization. B. Temperature optimization. C. The influence of distinct metal ions on CaUGT activity. D. The yield of Reb M at different reaction times. Error bars represent the standard deviation of three duplications. **Figure S11:** Analysis and Prediction of the Catalytic Mechanism for CaUGT.