nature communications

Article

https://doi.org/10.1038/s41467-025-57722-9

# Comparative genomics provides insights into the biogeographic and biochemical diversity of meliaceous species

Received: 11 September 2024

Accepted: 28 February 2025

Published online: 17 March 2025

Check for updates

Jia Liu[1,3], Zhennan Wang[1,3], Xinyao Su[1,3], Liang Leng[2], Jiarou Liu[1], Feng Zhang[1], Shilin Chen[1,2] ✉, Yujun Zhang[1] ✉ & Caixia Wang[1] ✉

Meliaceous plants such as *Azadirachta indica* (neem) and *Melia azedarach* (chinaberry) contain large amounts of limonoids with unique anti-insect activities. However, genes responsible for downstream modifications of limonoids are not well known. Here, we improve the genome assemblies of neem and chinaberry to the telomere-to-telomere (T2T) level. Allopatric speciation of the two plants is confirmed by the lineage-specific inversion of chromosome 12 in the neem lineage. We further identify two BAHD-acetyltransferases (ATs) in chinaberry (MaAT8824 and MaAT1704) that catalyse acetylation at both the C-12 and C-3 hydroxyl groups of limonoids, whereas the syntenic neem copy (AiAT0635) does not possess this activity. A critical N-terminal region (SAGAVP) is crucial for the acetylation of AiAT0635, and swapping it into the MaAT8824 version (CHRSSG) can endow it with acetylation activity. Our improved genome assemblies provide insights into allopatric speciation of neem, as well as limonoid biosynthesis and chemical diversity in meliaceous plants.

The name limonoid comes from limonin, the earliest tetranortriterpenoid obtained from citrus. As highly oxygenated and modified triterpenoids, these secondary metabolites are responsible for the bitterness of some Rutaceae and Meliaceae plants. Due to their unusual structural diversity, unique anti-insect activity and promising pharmaceutical value[1], limonoids have attracted much interest in recent decades, and approximately 2700 meliaceous limonoids have been identified to date[2]. Toosendanin is a tetranortriterpenoid extracted from chinaberry (*Melia azedarach* L.), which was used as a digestive tract-parasiticide and agricultural insecticide in ancient China, whereas azadirachtin is a seco-C-ring tetranortriterpenoid extracted from the neem tree (*Azadirachta indica* A. Juss). Both chinaberry and neem are Meliaceae plants. Chinaberry is a deciduous tree endemic to Southwest China, whereas the evergreen neem tree is thought to be endemic to the Indian subcontinent. The genomic divergence between neem and chinaberry and the biochemical diversity within and between the two plants encouraged the sequencing and assembly of their genomes[3–6]. Unfortunately, many unclosed gaps and assembly errors are still present in the latest versions, without haplotype-resolved telomere-to-telomere (T2T) assemblies available.

The biosynthesis of toosendanin and azadirachtin starts from 2,3-oxidosqualene, the common precursor derived from the mevalonate pathway, which was first cyclized into tirucalla-7,24-dien-3β-ol via the recently two identified oxidosqualene cyclases - AiOSC1 and MaOSC1[7]. The triterpene scaffold then undergoes more than 20 steps of modification, including oxidation, esterification, demethylation, side chain degradation, and ring opening. Among these steps, two cytochrome P450s (CYP450s), MaCYP71CD2 and MaCYP71BQ5, from chinaberry can sequentially oxidise tirucalla-7,24-dien-3β-ol into melianol[7]. The identification of 22 enzymes that catalyse sequential reactions from melianol to azadirone and kihadalactone A biosynthesis in Melia and

[1]State Key Laboratory for Quality Ensurance and Sustainable Use of Dao-di Herbs, Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, China. [2]Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu, Sichuan, China. [3]These authors contributed equally: Jia Liu, Zhennan Wang, Xinyao Su. ✉e-mail: slchen@icmm.ac.cn; yjzhang@icmm.ac.cn; cxwang@icmm.ac.cn

Citrus plants[8], respectively, represented the greatest breakthrough in the past decade.

Despite this progress, the elucidation of downstream modifications from azadirone to toosendanin and azadirachtin is still challenging. The biosynthesis of toosendanin and azadirachtin is speculated to diverge after azadirone or its analogues. For toosendanin, hydroxylation of C1, C11, C12, C14, C15, C19, C29, and acetylation of C3-OH and C12-OH of azadirone were proposed. For azadirachtin, further modifications involved the opening of the C ring, a series of oxidations, and intramolecular rearrangements[8]. Specifically, the difference in the acyl group on C12-OH suggests divergent functions of acetyltransferases between the two species.

In this work, we improve the genome assemblies of neem and chinaberry to the T2T level and identify two acetyltransferases (MaAT8824 and MaAT1704) catalysing the C12-OH acylation of limonoids. The syntenic neem copy (AiAT0635) is endowed with acylation activity by swapping six amino acids into the corresponding chinaberry version, which explains the chemical diversity between the two species (Fig. 1). Our study contributes to the understanding of limonoid biosynthesis and reveals the genetic basis of the biochemical diversity of Meliaceae.

## Results

### Assemblies and analysis of the *A. indica* and *M. azedarach* genomes

We used HiFiasm (v0.19.8) for the hybrid assembly of *A. indica* (neem) and *M. azedarach* (chinaberry) using HiFi, ONT and Hi-C data (Supplementary Table 1–3), obtaining 13 haplotype-resolved T2T chromosomes for each species. The last chromosome was revealed to be an rDNA-bearing one, which was neglected in previous assemblies[4,5,8]. Using droplet digital PCR, the copy numbers of rDNA in neem and chinaberry were estimated to be 551 and 159, respectively (Supplementary Table 4). We manually extended the rDNA-bearing chromosome to telomeres, resulting in a complete set of T2T assembly for each species. The per-base consensus quality values (QVs) exceeded 60 (Table 1), and BUSCO evaluation suggested that the genome completeness was greater than 99.5%, indicating significant improvements in completeness and continuity (Supplementary Table 5). We compared our T2T assemblies with the currently available neem and chinaberry genomes (Table 1), addressing the respective gaps within these genomes.

In the genomes of neem and chinaberry, repetitive sequences accounted for 32.14% and 32.10%, respectively, with 28,215 and 28,889 genes annotated (Supplementary Table 6–9). Phylogenetic analysis

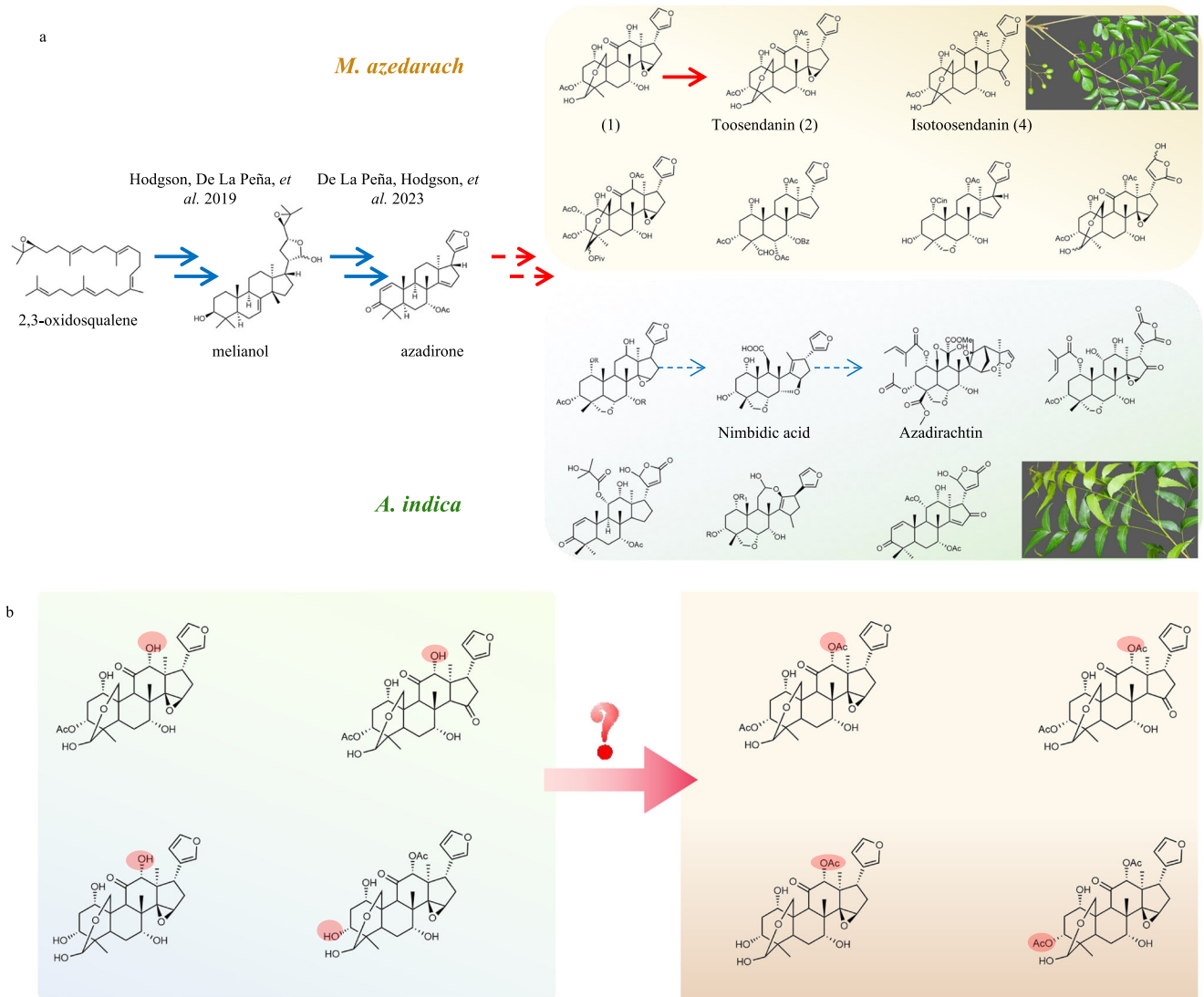


**Fig. 1 | Schematic representation of this work. a** Proposed limonoid biosynthetic pathway and its structural diversity in neem (*A. indica*) and chinaberry (*M. azedarach*). Specifically, neem limonoids are mainly C-ring open types, whereas chinaberry limonoids are usually acetylized towards C12-OH. **b** Representative limonoid substrates (bottom left) and their acetylated products in chinaberry (bottom right).

https://doi.org/10.1038/s41467-025-57722-9

**Table 1 | Comparison of assembly statistics of *A. indica* and *M. azedarach* genomes**

| Species | Sequencing platform | Assembly level | Genome size (Mb) | Repeat (%) | Gene Number | Base pair (QV) | Assigned (C) | BUSCO Comp. | Trans. map. | Reference |
|---|---|---|---|---|---|---|---|---|---|---|
| *Melia azedarach* | PacBio&ONT | T2T | 222.45 | 32.14 | 28889 | 66.40 | 100% | 99.60 | 93.56 | Present work |
| | PacBio | Chromosome | 230.8 | – | 26712 | – | – | 92.98 | – | 8 |
| | Illumina & ONT | Chromosome | 239.23 | 37.47 | 21983 | – | 99.14 | 96.05 | – | 5 |
| *Azadirachta indica* | PacBio& ONT | T2T | 233.21 | 32.10 | 28215 | 60.54 | 100% | 99.50 | 92.83 | Present work |
| | Illumina & ONT | Chromosome | 232.68 | 33.11 | 23087 | – | 96.13 | 99.07 | – | 5 |
| | PacBio | Chromosome | 281.70 | 40.99 | 25767 | – | 96.93 | 91.7 | – | 82 |
| | Illumina | Contig | 225 | 24.15 | – | – | – | – | – | 4 |
| | Illumina | Contig | 267 | 32.5 | 32278 | – | – | – | – | 83 |
| | Illumina | Contig | 261.10 | – | – | – | – | – | – | 83 |

QV = quality value, QV60 is no more than 1 base call error per 100,000 bp; C = percentage of the assembly assigned to chromosomes; Comp = Completeness; Trans. map. = Transcripts mappability.

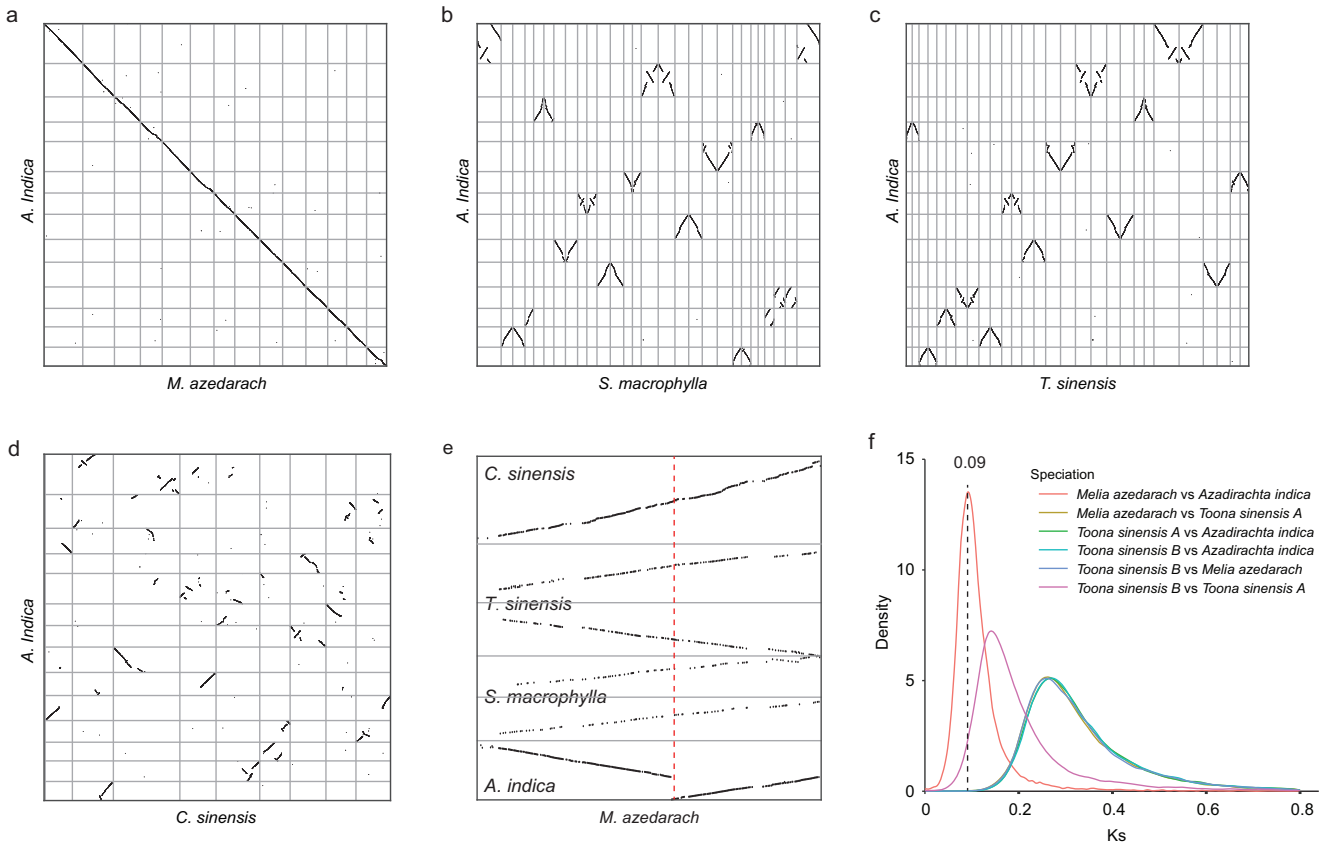


**Fig. 2 | Assembly and analysis of the neem and chinaberry genomes. a–d** Dot plots of the neem genome compared with four species, including *M. azedarach* (**a**), *S. macrophylla* (**b**), *T. sinensis* (**c**), and *C. sinensis* (**d**). **e** Enlarged dot plots of chinaberry compared with four species towards the long arm end of Chr12. Lineage-specific inversion on the neem chromosome is obvious. **f** Distribution of Ks for paired syntenic genes between three meliaceous species. The tetraploid *T. sinensis* genome was arbitrarily split into two subgenomes for comparison.

using 138 single-copy 1:1:1 orthologous genes across 19 species supported two clades of Sapindales. Clade one included Meliaceae, Rutaceae, and Anacardiaceae, whereas clade two included Sapindaceae. Within Meliaceae, a chromosomal comparison revealed high-level synteny between neem and chinaberry, whereas both *Toona sinensis* and *Swietenia macrophylla* underwent whole-genome duplications (Fig. 2a–d).

## Karyotypic evolution underlines allopatric speciation of *A. indica*

While the evergreen neem trees are widely observed from Africa to America, they are endemic to Assam of the Indian subcontinent, and their global distribution started approximately 100 years ago through private or commercial initiatives[9]. On the other hand, chinaberry is endemic to the Sichuan Province of Southwest China. The distance between the two endemic sites is approximately 700 miles, passing through the Himalaya–Hengduan Mountains. The vertical distribution of neem can reach 1500 m, while most mountains of the Himalaya–Hengduan region are higher than 4000 m, preventing further gene flow between the two sites. This implied that the distinct morphological traits of the two species resulted from allopatric speciation (Supplementary Fig. 1).

We performed a comparative analysis of the two genomes to evaluate the genetic basis of species divergence. Corresponding

chromosomes between neem and chinaberry are highly conserved and syntenic, with only one terminal inversion of 1.08 Mb on chr 12 observed. Detailed comparison with outgroup species of Sapindales confirmed that this inversion occurred only in the neem lineage (Fig. 2e), suggesting that chinaberry represents the ancestral chromosomal block while neem is derivative.

Chromosomal inversion has long been viewed as driving force of speciation, which will disrupt recombination in heterozygotes by reducing crossing-over within inverted regions[10–13]. Gene flow will then be restricted, leading to final speciation. It seemed that the neem-chinaberry species pair, which were isolated by uplift of the Himalaya–Hengduan Mountains, represents an ideal model of allopatric speciation.

We calculated the divergence of orthologous syntenic gene pairs between neem and chinaberry ($n$ = 18,165), which peaked at approximately 0.09 (Fig. 2f). The neutral mutation rate of $2.5 \times 10^{-9}$ mutations per bp per year[14] implies that allopatric speciation of neem from chinaberry started approximately 18 MYA. Till now, the rise and growth history of the Tibetan Plateau has been extensively analysed using geological data, with various estimates ranging from 15 to 55 MYA[15,16]. The allopatric speciation of neem from chinaberry provided a rare biological perspective on the geological evolution of the Tibetan Plateau, which agreed well with the cessation time of rapid Pacific trench migration (15–20 MYA)[17].

## Expansion of genes involved in the biosynthesis of sulphur-containing volatiles in Meliaceae

Meliaceae trees, notably those of the *Toona* and *Azadirachta* species, have sulphur-containing volatiles with garlic-like smells[18]. Tender leaves of *T. sinensis*, which contain abundant diallyl disulphide derived from γ-glutamyl-*S*-ally-L-cysteine, have been enjoyed by the Chinese as woody vegetables for thousands of years. The proposed biosynthetic pathway of volatile organosulphur compounds includes (1) the conjugation of cysteine and glutamic acid by γ-glutamylcysteine synthetase, (2) the addition of glycine to the C-terminal site of γ-glutamylcysteine by glutathione synthetase, (3) the *S*-conjugation of glutathione, (4) the removal of the glycyl group by phytochelatin synthase, (5) the modification of the *S*-alk(en)yl group, and (6) the removal of the γ-glutamyl group by γ-glutamyl transpeptidase (GGT)[19] (Supplemental Fig. 2).

Accumulation of ordor compounds implies possible expansion of genes involved in sulphur-containing volatile biosynthesis in Meliaceae. As shown in Fig. 3, *C. sinensis* has only one GGT copy on chr1, and a total of 16 and 14 copies were observed on the orthologous neem and chinaberry chromosomes, respectively (Fig. 3a, b). A significant expansion of GGTs was also detected in the *T. sinensis* genome, but the tetraploid *S. macrophylla* genome had only one copy, suggesting that this expansion occurred during Meliaceae species radiation. In comparison, the garlic genome is characterised by extensive duplication of downstream alliinase genes, which catalyse the hydrolysis of *S*-alk(en)yl-l-cysteine sulfoxides[20]. High expression of tandemly duplicated GGTs, particularly in stems and leaves, was observed, which presumably contributed to the distinctive odours of these plants (Fig. 3c). We quantified the content of *S*-allylcysteine in the leaves of these species. It turned out that the content sorting was *T. sinensis* > neem > chinaberry > *C. sinensis*, which was consistent with GGT copies in these species (Fig. 3b and Supplementary Fig. 2).

## Expansion of acyltransferases for limonoid biosynthesis in Meliaceae

Meliaceous plants contain large amounts of limonoids with unique anti-insect activities[21–23]. Despite recent progress in protolimonoid biosynthesis, the elucidation of the complete biosynthetic route of diverse limonoids from azadirone to toosendanin or azadirachtin remains a major challenge. Hydroxylation and acetylation, which are catalysed by CYP450s and ATs, respectively, are involved in protolimonoid modifications[8,24] and limonoid diversity. For this reason, we focused on acyltransferase (AT) family genes in neem and chinaberry.

Systematic annotation and curation revealed 11 ATs in *C. sinensis* and various copies in Meliaceae plants, from 15 in *T. sinensis* to 39 in chinaberry. Tandem clusters of 9 and 15 ATs on chr3 were identified in neem and chinaberry, respectively. Meanwhile, 8, 6, and 0 copies were annotated in syntenic regions of outgroup species *T. sinensis, S. macrophylla* and *C. sinensis*, respectively (Fig. 3d), suggesting that tandem expansion postdates Meliaceae species radiation but predates the divergence of neem and chinaberry.

We performed transcriptome sequencing of different neem and chinaberry tissues in triplicate. Weighted gene co-expression network analysis (WGCNA) revealed a gene module that strongly correlated with fruit in chinaberry, containing 1352 genes. The GO enrichment analysis of these genes revealed that among the top five enriched GO terms, two were significantly associated with acetylation functions: acyltransferase activity and O-acyltransferase activity (Supplementary Fig. 3). We subsequently extracted all the genes from the two GO terms and aligned them with all the chinaberry proteins, then the tandem cluster of ATs on chr3 stood out. We further analysed the expression of these ATs. While the syntenic relationships of these ATs were obvious, their expression patterns diverged extensively, especially in fruits (Fig. 3e). We postulated that different expression levels of these ATs in fruits contributed to acetylation diversity of limonoids between neem and chinaberry.

## Characterisation of the ATs for limonoid C12-OH acylation

Since acylation at C12 is a divergent feature between limonoids of neem and chinaberry which present different activities against insects and cancer cell lines[21–23,25]. We speculate that the AT gene cluster from WGCNA of chinaberry might be involved. We selected six chinaberry ATs (*MaAT-1, 3, 6, 7, 9*, and *15*) and four neem copies (*AiAT-1, 2, 3*, and *4*) from the tandem clusters for in vitro functional characterisation using *Escherichia coli*, since these ATs are highly expressed in fruit, as indicated by WGCNA (Supplementary Fig. 3). First, all the candidate *MaATs* were expressed in *E. coli*. Two compounds **1** and **3** were then prepared by chemical hydrolysis, and their structures were further verified by nuclear magnetic resonance (NMR) spectroscopy (Supplementary Figs. 4 and 5). The recombinant proteins were assessed with compound **1** as the acceptor and acetyl-CoA as the acyl donor. The liquid chromatography–mass spectrometry (LC-MS) analysis revealed that the MaAT8824 and MaAT1704 proteins formed a product with the same retention time (RT = 5.48 min) and mass spectrum ($[M+Na]^+$ = 597.2215) as authentic toosendanin **2** (Fig. 4a–d and Supplementary Fig. 6). When compound **3** was used as the substrate, isotoosendanin **4** was detected (Fig. 4a, e–g and Supplementary Fig. 7). It's noteworthy that the additional peaks at 6.5 min in Fig. 4b and at 4.3 min in Fig. 4e correspond to isomers of compounds **2** and **4**, respectively, due to the existence of the unstable hemiacetal hydroxyl group in their structures[26]. Therefore, MaAT-6 (MaAT8824) and MaAT-7 (MaAT1704) can transfer acetyl groups into the C12-hydroxyl moieties of compounds **1** and **3** and produce toosendanin **2** and isotoosendanin **4**, respectively. The remaining four chinaberry ATs (MaAT-1, 3, 9, and 15) failed to acylate compounds **1** and **3** (Supplementary Fig. 8). We also characterised the function of MaAT8824 and MaAT1704 in *Nicotiana benthamiana*. LC-MS analysis of *N. benthamiana* leaf extracts confirmed the acetylation of compounds **1** and **3** into compounds **2** and **4** towards C12-OH, respectively, which agreed well with results from *E. coli* (Supplementary Fig. 9).

We quantified the contents of compounds **1** to **4** in different tissues. Compounds **1** and **3** had the highest contents in the fruits of both plants (Fig. 4h, j), whereas compounds **2** and **4** were detected only in chinaberry, with the highest contents in fruits (Fig. 4i, k). This finding

https://doi.org/10.1038/s41467-025-57722-9

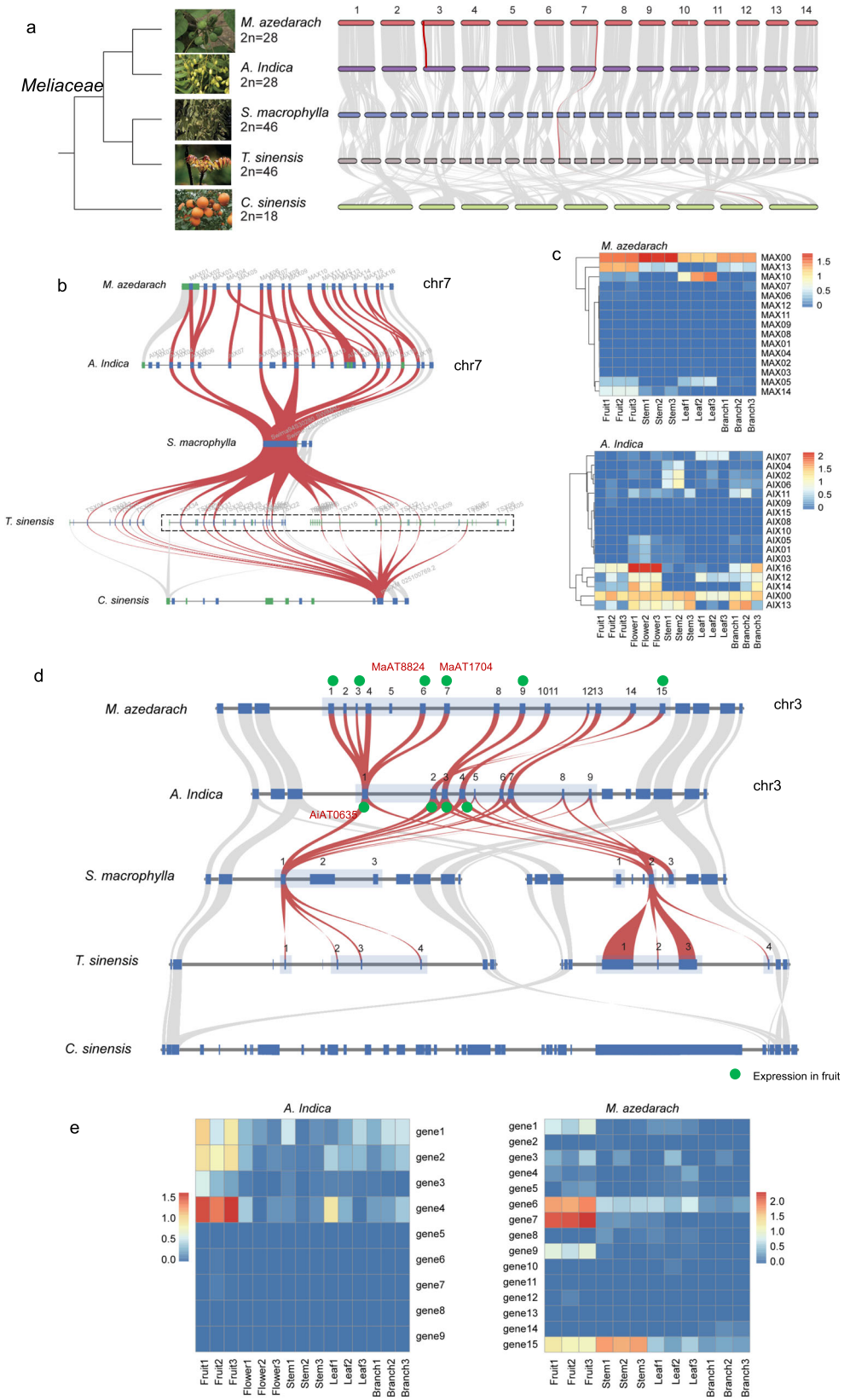


**Fig. 3 | Lineage-specific tandem duplications of GGT and AT genes. a** Phylogeny of five Sapindales species with pairwise chromosome alignments. Each horizontal bar represents one chromosome, and the chromosome IDs of the chinaberry are labelled at the top. Chromosomes that experienced GGT duplication are highlighted in colour. **b** Orthologous relationships of duplicated GGTs across five species. The synteny of genes around GGTs is illustrated by grey bands connecting orthologues across species. **c** Heatmaps of GGT expression across different tissues. The expression of each gene was measured as a log (1 + TPM). **d** Orthologous relationships of duplicated ATs across five species. **e** Heatmaps of AT expression across different tissues.

https://doi.org/10.1038/s41467-025-57722-9

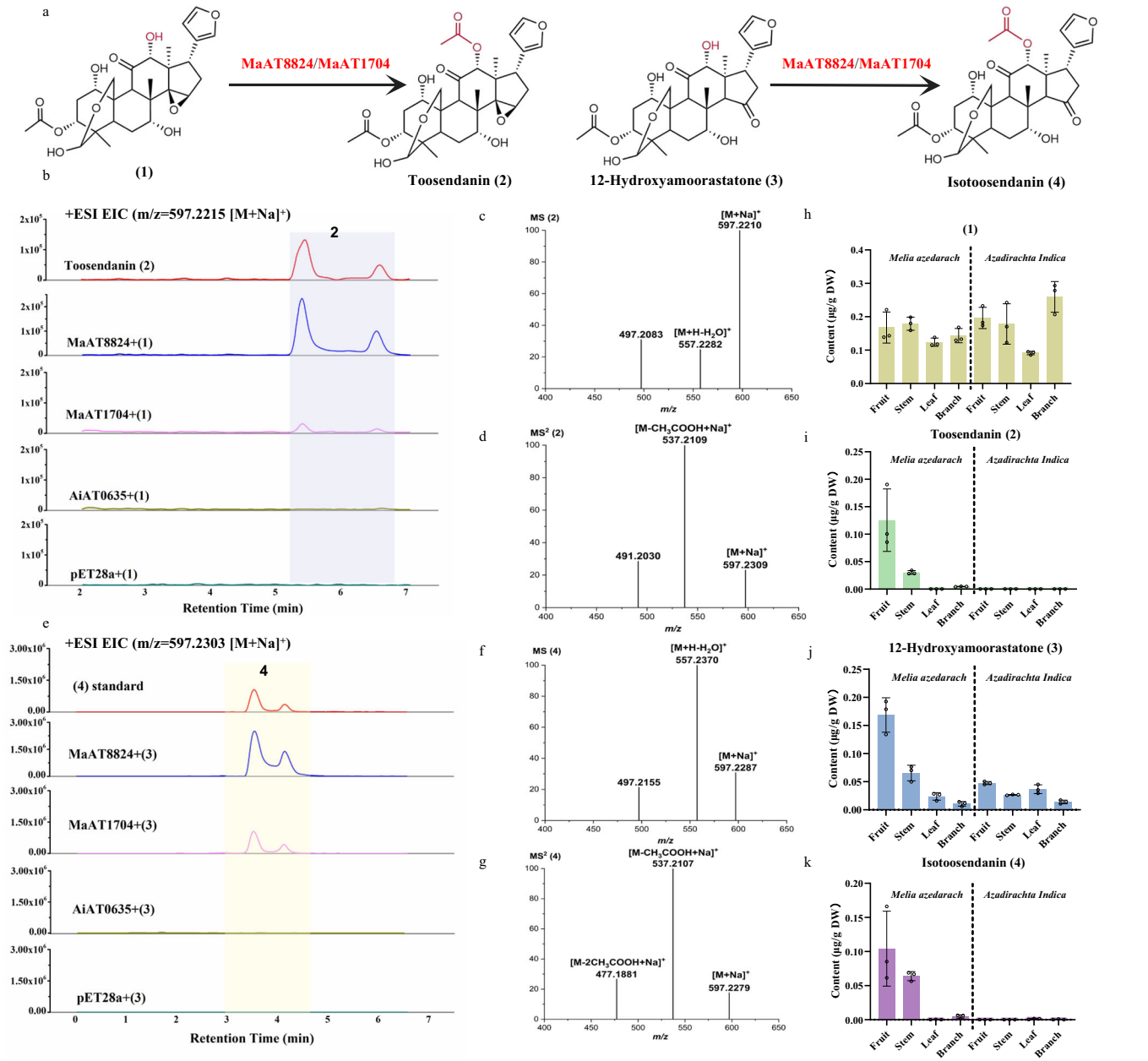


**Fig. 4 | Identification of ATs that catalyse acylation in toosendanin/isotoosendanin biosynthesis. a** Reactions for compounds **1** to **2** and compounds **3** to **4** catalysed by MaAT8824 and MaAT1704. **b** LC-MS chromatograms of the standard, the control samples and the reaction mixtures for MaAT8824 and MaAT1704 using acetyl-CoA as the acetyl donor and compound **1** as the substrate. **c**, **d** LC-MS and LC-MS/MS analyses of reaction product **2** for MaAT8824 and MaAT1704 in positive ion mode; **e** LC-MS chromatograms of the standards, the control samples and the reaction mixtures for MaAT8824 and MaAT1704 using acetyl-CoA as the acetyl donor and compound **3** as the substrate. **f**, **g** LC-MS and LC-MS/MS analyses of reaction product **4** for MaAT8824 and MaAT1704 in positive ion mode. **h**–**k** The contents of the four compounds in the four tissues of the two plants. The data are presented as the means ± SDs ($n = 3$ biological replicates). Source data are provided as a Source Data file.

agreed well with the transcriptomic and enzymatic results. These observations suggested that MaAT8824 and MaAT1704 could transfer an acetyl group into the C12-OH of limonoids to finalise toosendanin/isotoosendanin biosynthesis in chinaberry. In contrast, the syntenic neem genes have no acetylation activity, resulting in limonoid diversity between the two allopatric species.

Missing of compounds **2** and **4** in neem may result from the loss of function of the syntenic neem genes or neo-functionalization of the chinaberry copies[27]. To clarify the exact evolutionary scenario, genes from neem and outgroup species were assayed for acetylation activity. As expected, none of the neem copies could transfer acetyl groups to C12-OH of compounds **1** or **3** (Fig. 4b, e and Supplementary Fig. 8). Four ATs from *S. macrophylla*, *T. sinensis*, and *C. Sinensis* were selected according to collinear relationship (Fig. 3d and Supplementary Fig. 10), and MaAT8824 was used as positive control. These genes were expressed in *E. coli*, and the recombinant proteins were assessed with compound **1** as the acceptor and acetyl-CoA as the acyl donor. Results suggested that all these ATs, except MaAT8824, had no acetylation activity towards compound **1** (Supplementary Fig. 11). Accordingly, we propose that the C12-OH acylation activities of MaAT8824/MaAT1704 represented neo-functionalization and emerged after divergence of chinaberry and neem.

https://doi.org/10.1038/s41467-025-57722-9

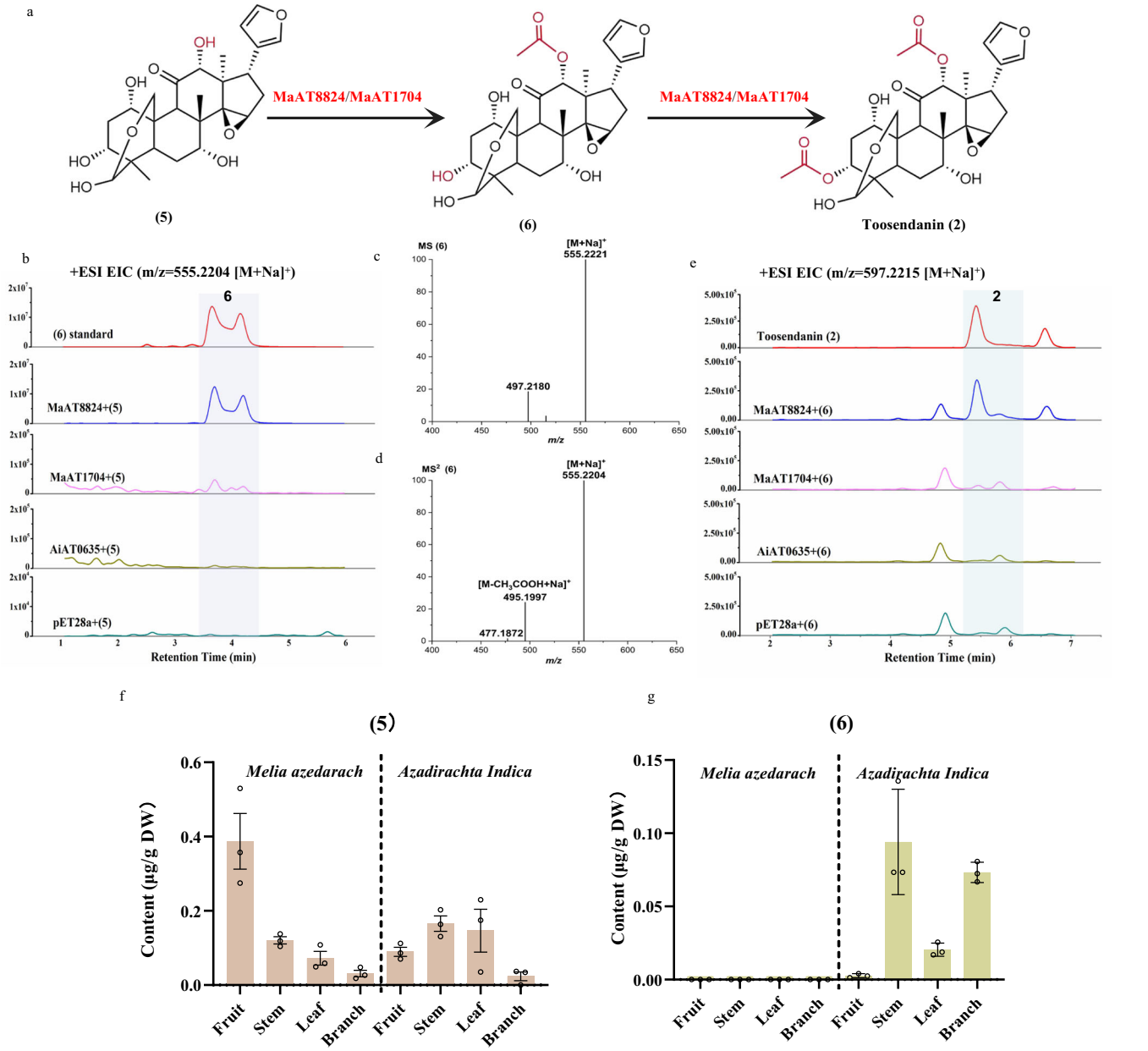


**Fig. 5 | Acylation activity of chinaberry ATs towards the C3-OH group of limonoid. a** Reactions from compounds **5** to **6** and compounds **6** to **2** catalysed by MaAT8824 and MaAT1704. **b** LC-MS chromatograms of the standard samples, the control samples and the reaction mixtures of MaAT8824 and MaAT1704 with acetyl-CoA as the acetyl donor and compound **5** as the substrate. **c**, **d** LC-MS and LC-MS/MS analyses of the reaction product **6** catalysed by MaAT8824 and MaAT1704 in positive ion mode. **e** LC-MS chromatograms of the reaction mixtures for MaAT8824 and MaAT1704 using acetyl-CoA as the acetyl donor and compound **6** as the substrate. **f**, **g** The contents of compounds **5** and **6** in different neem and chinaberry tissues. The data are presented as the means ± SDs ($n = 3$ biological replicates). Source data are provided as a Source Data file.

## Acylation activity of MaAT8824 towards the C3-OH group of limonoid 6

We then tested the acylation activity of MaAT8824/MaAT1704 towards other positions, such as C3-OH, by preparing compounds **5** and **6** (Supplementary Figs. 12 and 13). When compound **5** was used as the substrate, to the best of our knowledge, a previously unobserved product was detected for MaAT8824 and MaAT1704. The product was confirmed to be compound **6** when the retention time (RT time = 3.6 min) and mass spectrum ($[M+Na]^+ = 555.2204$) were compared with those of the authentic product **6** (Fig. 5a–d and Supplementary Fig. 14). This result indicated that MaAT8824 and MaAT1704 could also catalyse acetylation at the C-12 hydroxyl group of compound **5** to form compound **6**. Interestingly, compound **6** can be further catalysed into toosendanin **2** by C3-OH acylation (Fig. 5a, e and Supplementary Fig. 15).

The acylation activity of MaAT8824/MaAT1704 against compound **14**, which was derived by deacetylation of compound **4** at both C-3 and C-12 (Supplementary Fig. 16), was further tested. While a peak of suspected compound **15** was detected using MaAT8824 (Supplementary Fig. 17), attempts to purify this compound failed because of low catalytic efficiency (Supplementary Fig. 18), and the catalytic capacity was even lower when using MaAT1704 (Supplementary

https://doi.org/10.1038/s41467-025-57722-9

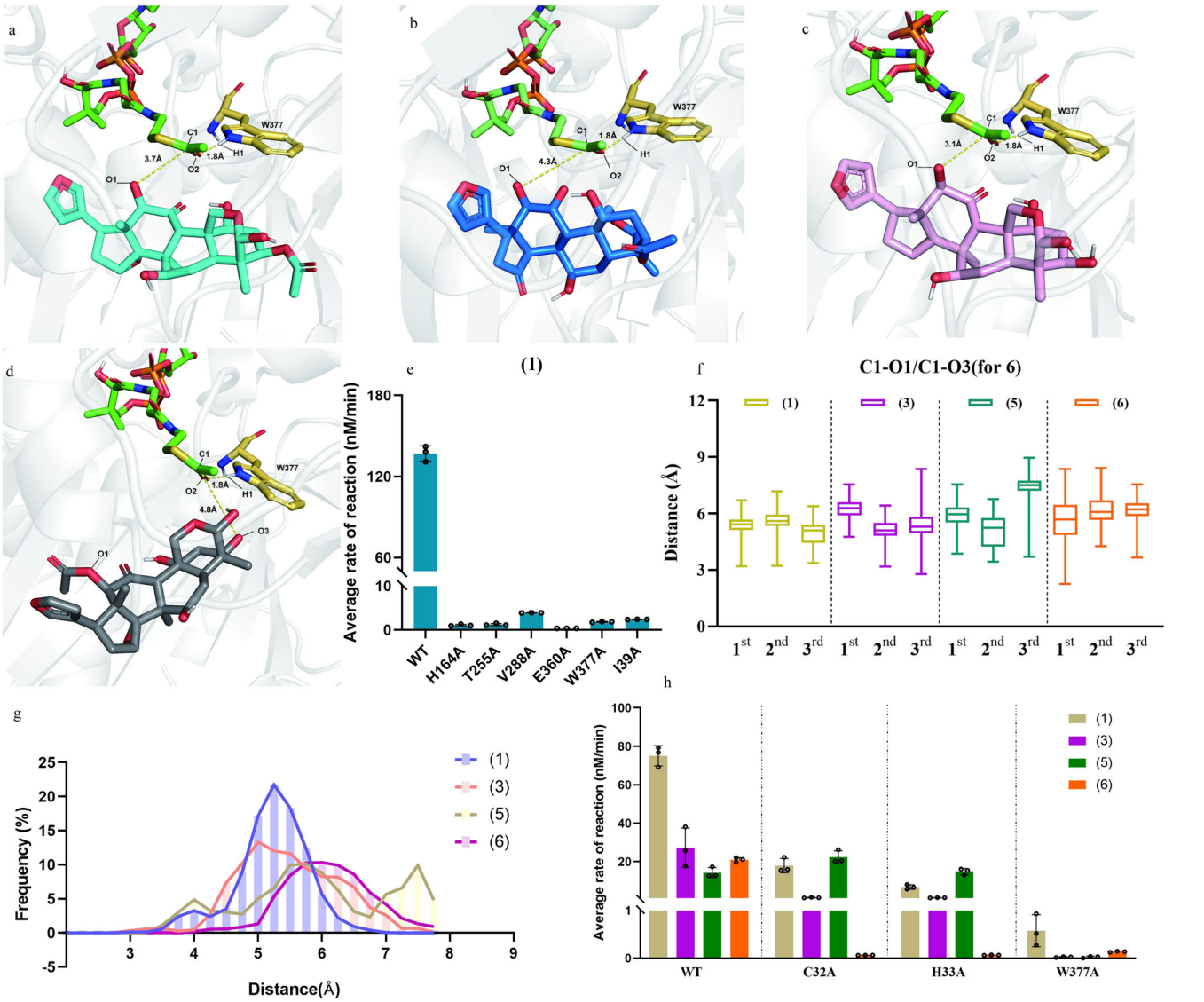


**Fig. 6 | Structural basis of the MaAT8824 catalytic preference for C12-OH groups. a–d** Docking conformations of compounds **1**, **3**, **5**, and **6** and Ac-CoA into MaAT8824. **e** In vitro assay of the MaAT8824 mutants using compound **1** as the substrate. **f** The distance of C1-O1/C1-O3 during the MD simulations for compounds **1**, **3**, **5**, and **6** (C1-O1 for compounds **1**, **3**, and **5**, and C1-O3 for compound **6**). Box plots show median values (solid horizontal lines), 25th and 75th percentile values (box), and 90th percentile values (whiskers). **g** The distance (C1-O1/C1-O3) frequency during the MD simulations of compounds **1**, **3**, **5**, and **6**. **h** The average rates of reaction of three MaAT8824 mutants for compounds **1**, **3**, **5**, and **6**. The data are presented as the means ± SDs ($n = 3$ biological replicates). Source data are provided as a Source Data file.

Fig. 17). As expected, the remaining ATs within the chr3 cluster had no acetylation activity towards C3-OH.

Taken together, these results implied that MaAT8824 and MaAT1704 could catalyse the acetylation of both the C-12 and C-3 hydroxyl groups of limonoids, with a preference for the C-12 hydroxyl group. These activities were also confirmed in vivo in *N. benthamiana* (Supplementary Fig. 19).

It's noteworthy that high contents of compound **5** were detected in both chinaberry and neem, especially in fruits, yet compound **6** was only detected in neem (Fig. 5f, g). This may be attributed to the higher catalytic activity of MaAT8824 towards compound **6** than compound **5**, resulting in rapid consumption of **6**. On the other hand, since the syntenic neem ATs had no acylation activity, accumulation of compound **6** in neem might imply the existence of unidentified AT gene(s) from elsewhere of the genome with C12-OH acylation activity towards **5**, which warrants further analysis.

## Catalytic mechanism of MaAT8824

Since MaAT8824 possesses higher catalytic activity than MaAT1704 (Supplementary Fig. 20), attempts were made to obtain the complex crystal structure of MaAT8824 but failed. We therefore predicted its structure using alphafold2 with PLDDT = 90. We performed molecular docking with compound **1** as the acceptor and acetyl-CoA as the donor into the catalytic pocket, aiming to elucidate the catalytic mechanism of MaAT8824. Ac-CoA and compound **1** were docked into the structure according to the donor and substrate positions in SbHCT[28] and AmAT7-3[29] (Fig. 6a and Supplementary Fig. 21). Compounds **3**, **5** and **6** were also docked in the same way (Fig. 6b–d). We selected key residues within 5 Å that interact with the substrate or donor and mutated them according to the docking results with compound **1** to verify the reliability of our analysis. Mutation of residues I39, H164, E360, and W377 to alanine led to enzyme inactivation, confirming the reliability of the molecular docking data (Fig. 6e).

**Table 2 | Kinetic parameters of MaAT8824 for compounds 1, 3, 5 and 6**

| Substrate | Acyl donor | Acylation site | *Km* [μM] | *kcat* [$s^{-1}$] | *kcat/Km* [$s^{-1}$ $M^{-1}$] |
|---|---|---|---|---|---|
| **1** | Acetyl-CoA | 12-OH | 64.1 | 0.020 | 318.43 |
| **3** | Acetyl-CoA | 12-OH | 205.1 | 0.037 | 180.2 |
| **5** | Acetyl-CoA | 12-OH | 390.7 | 0.014 | 35.1 |
| **6** | Acetyl-CoA | 3-OH | 232.5 | 0.024 | 106.8 |

Molecular dynamics (MD) simulations were further conducted with three replicates to understand the acetylation mechanism of this enzyme at C12-OH and C3-OH. In the acetylation of BAHD-AT, the nucleophilic attack of the oxyanion of substrates towards the carbonyl carbon of the acyl donor is critical for catalysis[28]. Therefore, the distances between the acyl C and the substates, C1-O1/C1-O3 (C1-O1 for compounds **1**, **3**, and **5**, and C1-O3 for compound **6**), were calculated elaborately. The average distances between the acyl C and the oxyanions in compounds **1** and **3** are 5.34 Å and 5.6 Å, respectively, which are shorter than those in compounds **5** and **6** (> 6.1 Å) (Fig. 6f). The distance frequencies between the acyl C and the oxyanion in the four substrates are shown in Fig. 6g. For compounds **1** and **3**, this distance was predominantly within 5 Å but was mostly longer than 6 Å for compounds **5** and **6**. These simulations indicated that the relatively shorter distance between the acyl C and the oxyanion of compound **1** and the steady binding of Ac-CoA are responsible for its preferential catalysis toward C12-OH.

The substrate specificity of MaAT8824 was further validated by kinetic analysis. The apparent affinities ($K_m$) of MaAT8824 for acceptors **1**, **3**, **5**, and **6** were calculated as 64.06 μM, 205.1 μM, 390.7 μM, and 232.5 μM, respectively. Moreover, the catalytic efficiencies ($k_{cat}/K_m$) of MaAT8824 with respect to compounds **1**, **3**, **5**, and **6** were 318.4 $s^{-1}$ $M^{-1}$, 180.2 $s^{-1}$ $M^{-1}$, 35.1 $s^{-1}$ $M^{-1}$, and 106.8 $s^{-1}$ $M^{-1}$, respectively. These results implied that the catalytic efficiency of MaAT8824 towards compound **1** was much greater than that towards compounds **5** and **6** (Table 2 and Supplementary Fig. 22). A comparison of compounds **1**, **6** and **5** implied that MaAT8824 presented higher catalytic activity towards substrates with acetyl groups than towards those without acetyl groups, regardless of the presence of C3-OH or C12-OH. Finally, the preference for substrate with C12-OH rather than C3-OH was evident by comparing compounds **1** and **6**. Taken together, the substrate preferences of MaAT8824 were in the order of compounds **1** > **3** > **6** > **5**, which was consistent with the MD analysis. Kinetic analysis of MaAT1704 was also conducted, and the data indicated that the substrate preference of MaAT1704 was in the order of compounds **1** > **3** > **5** > **6** (Supplementary Fig. 22). These data revealed that both MaAT8824 and MaAT1704 preferentially catalyse the acetylation of the C12-OH of compounds **1** and **3**, and MaAT8824 presented notable higher catalytic activity.

We further conducted site-specific mutagenesis to elucidate the structural basis of substrate specificity. To this end, dozens of key amino acids located in the catalytic pocket were mutated, and the substrate specificity of these mutants was detected by calculating their average rates of reaction towards compounds **1**, **3**, **5**, and **6** (Fig. 6h). Among these mutants, C32A decreased reaction towards compound **1**, **3**, **6** and increased reaction towards compound **5**, indicating that the C32 residue plays a critical role in substrate specificity. Similarly, compared to the wild type, H33A decreased reaction rates by 90% towards compounds **1**, **3** and **6**, but had no obvious effect towards compound **5**. Moreover, W377A cannot catalyse compounds **3** and **5**, with residual activity against compounds **1** and **6**, which was consistent with the MD simulation results (Fig. 6h).

In addition to compounds **1**, **3**, **5**, and **6**, other compounds, including compounds **7**, **9**, and **13**, were also tested as possible acetyl acceptors of MaAT8824 (Supplementary Fig. 23). However, no products were observed when these chemicals were used as substrates (Supplementary Fig. 23). Different acyl donors, including malonyl-CoA, isobutyryl coenzyme A, succinyl coenzyme A, and benzoyl-CoA, were tested using compound **1** as the acceptor. MaAT8824 could not recognise these acceptors but could recognize only acetyl coenzyme A (Supplementary Fig. 24). Taken together, these results implied that MaAT8824 has relatively strict substrate specificity towards donors and slightly broader substrate specificity towards acceptors.

### A critical region drives the gain-of-function of AiAT0635

Neo-functionalization of MaAT8824 and MaAT1704 for C12-OH acetylation of limonoids, and the existence of non-functional ancestral copy in neem (AiAT0635), enable us to identify critical regions responsible for emergence of unique function. We aligned the three protein sequences to zoom in candidate residues (Supplementary Fig. 25). We observed seven divergent regions along the whole alignment, among which a critical 7-aa region towards the N-terminus (31–37) diverged significantly between MaAT8824 and AiAT0635 (Fig. 7a and Supplementary Fig. 26). We generated seven mutants by substituting the seven divergent regions of AiAT0635 into the corresponding sequence of MaAT8824 (Supplementary Fig. 26a), after which the catalytic activity of these mutants was assayed in vitro using compound **1** as a substrate. Interestingly, a peak appeared when AiAT0635-8824-F1 was overexpressed (substituting segment 31–37 (ISAGAVP) of AiAT0635 into segment 31–37 (LCHRSSG) of MaAT8824). The distinct product was proven to be toosendanin **2** (Supplementary Fig. 26b). Compared with that of MaAT8824, the conversion rate of this mutant was recovered to approximately 7.3%, even higher that of wild MaAT1704 (Fig. 7b). In comparison, no product was detected for the remaining six mutants (Supplementary Fig. 26). The acetylation activity of AiAT0635-8824-F1 was also confirmed in *N. benthamiana* (Supplementary Fig. 27). Conversely, we swapped the segment of MaAT8824 into those of MaAT1704 (ISASAAP) and AiAT0635 (ISAGAVP). Compared with that of MaAT8824, the relative conversion rates decreased by 15% and 60% for MaAT8824-mut-1704 and MaAT8824-mut-0635, respectively (Fig. 7b). The acetylation activity of MaAT1704 was completely abolished when its segment (ISASAAP) was swapped into the AiAT0635 version (Fig. 7b). Taken together, these results confirmed the critical role of the segment (LCHRSSG) for optimal acetylation activity in MaAT8824, while the segment (ISAGAVP) of AiAT0635 present no acetylation activity. The critical region was further narrowed down to segment 32–37 (SAGAVP) by substituting individual amino acids in different combinations (Fig. 7c–e and Supplementary Fig. 26c). These results suggested that the acylation function of the neem copy could be resurrected by substitution of six amino acids within the critical region.

We performed MD simulations of AiAT0635 and AiAT0635-mut with compound **1** as the acceptor and acetyl-CoA as the donor to elucidate the possible catalytic mechanism of AiAT0635-mut. Compound **1** was docked into AiAT0635 and AiAT0635-mut based on the two key distances between the O atom of the acetyl donor and the H atom of the catalytic tryptophan (O2-W377) and between the C atom of acetyl-CoA and the O atom of the acceptor substrate (C1-O1). We found that the relative position of compound **1** was overturned in AiAT0635-mut since the mutated amino acids have larger side chains (such as 32 C), leading to a smaller cavity and a changed position of the substrate (Fig. 7f, g and Supplementary Fig. 28a, b). According to the

https://doi.org/10.1038/s41467-025-57722-9

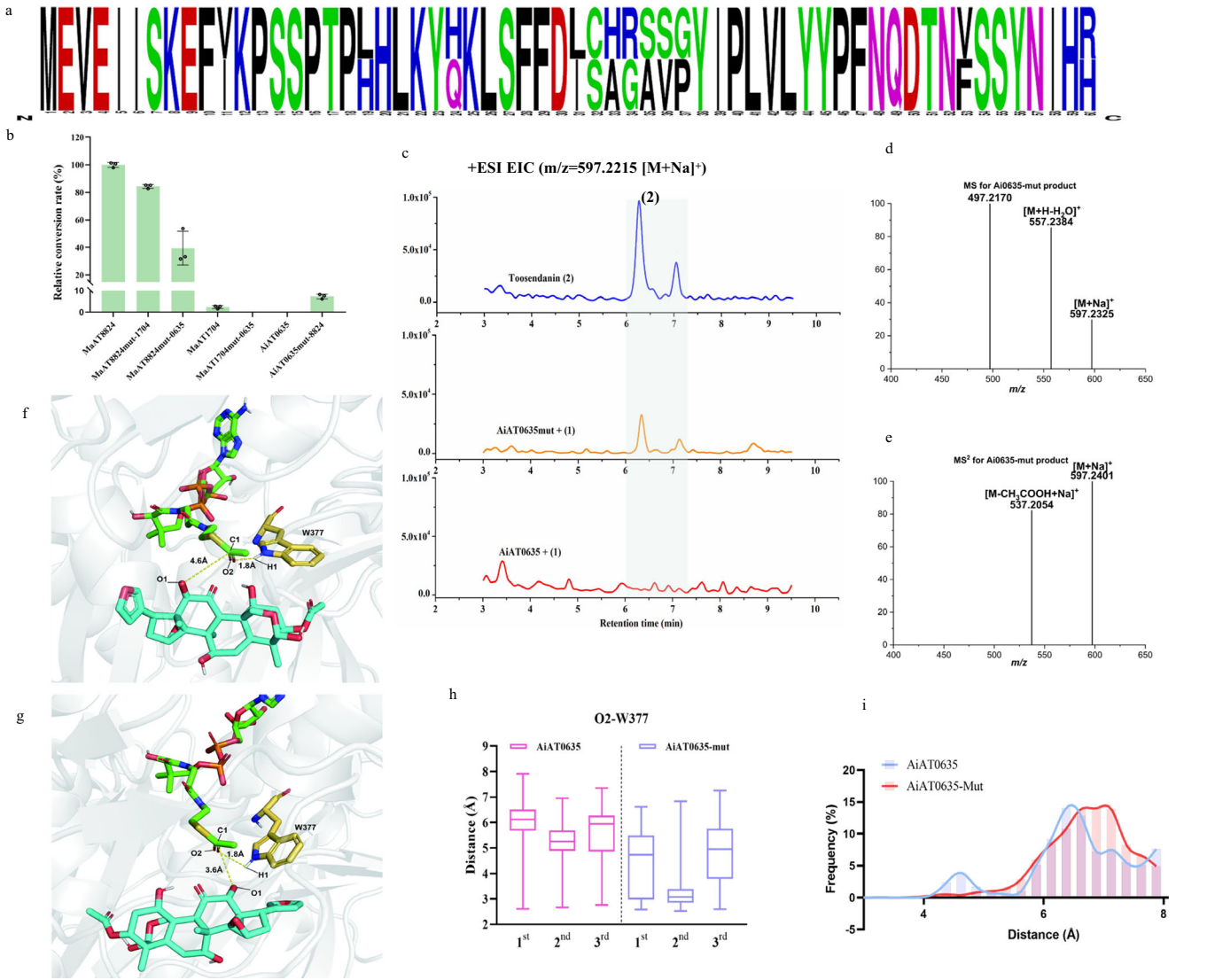


**Fig. 7 | A critical region drives the gain-of-function of AiAT0635. a** N-terminal sequence alignment of MaAT8824 and AiAT0635. **b** Relative conversion rates of different mutants when compound **1** was used as the substrate. **c** LC/MS chromatograms of the standard and the reaction mixtures for AiAT0635 and AiAT0635-mut using compound **1** as the substrate and acetyl-CoA as the acetyl donor. **d**, **e** LC-MS and LC-MS/MS analyses of products detected in reaction mixtures for AiAT0635-mut in positive ion mode. **f**, **g** Reactive snapshots of conformations for compound **1** in AiAT0635 and AiAT0635-mut from MD simulations. **h** The distance (O2-W377) between acyl O and W377 during the MD simulations using Substrate **1**. Box plots show median values (solid horizontal lines), 25th and 75th percentile values (box), and 90th percentile values (whiskers). **i** The distance (C1-O1/C1-O3) frequency in AiAT0635 and AiAT0635-mut. The data are presented as the means ± SDs ($n = 3$ biological replicates). Source data provided as a Source Data file.

results of the MD simulations, the root mean square deviations (RMSDs) of the two proteins and the donor did not differ substantially, whereas the RMSD of acetyl-CoA in AiAT0635-mut varied slightly after 6 ns, implying steady binding of the acetyl donor in AiAT0635-mut compared with that in AiAT0635 (Supplementary Fig. 28c–h). Moreover, the key distance of O2-W377 for AiAT0635-mut was 4.17 Å, which was shorter than that of wild-type AiAT0635 (5.6 Å) (Fig. 7h). The steady binding of the acetyl donor here was also consistent with the RSMD results. Taken together, these results indicate that the stable binding of the acetyl group of the donor in the protein may account for the gain-of-function of AiAT0635 after mutation. Another key distance of C1-O1 for AiAT0635-mut was similar to that of AiAT0635 (Fig. 7i and Supplementary Fig. 29), indicating that the distance between the donor and substrate did not change substantially after mutation. This result may explain why although AiAT0635-mut exhibited acetylation activity, its catalytic activity was still lower than that of MaAT8824 (Fig. 7b).

## Discussion

The biosynthesis of meliaceous limonoids has garnered particular attention owing to their wide range of biological activities and complex scaffold modifications. While several versions of genome assemblies for meliaceous species have been published, their completeness and contiguity are fairly low. In this study, we improved the assemblies of neem and chinaberry with haplotype-resolved T2T genomes and conducted comparative analyses to dissect the genetic basis of their biogeographical and biochemical diversity.

Neem and chinaberry are close relatives with highly conserved genome synteny, with only one chromosomal inversion observed. A comparison with outgroup species revealed a neem-specific inversion of 1.08 Mb on chr12. Interestingly, these two plants are endemic to the Indian subcontinent and Southwest China, respectively. Allopatric speciation of neem from chinaberry should have accelerated by the uplift of the Himalaya–Hengduan Mountains, which prevented further gene flow between the two plants. The estimated divergence time of 18

MYA agreed well with the cessation time of rapid Pacific trench migration, providing a rare biological perspective on the geological evolution of the Tibetan Plateau.

One of the differences between the limonoids of neem and chinaberry is their acetylation patterns. In this study, two ATs (MaAT8824 and MaAT1704) were identified to transfer acyl groups to different positions of limonoids (preferentially C12-OH) and were involved in toosendanin/isotoosendanin biosynthesis in chinaberry. While the neem syntenic copy did not present this acetylation activity, the swap of six amino acids (SAGAVP) into that of the chinaberry version (CHRSSG) endowed the activity, providing molecular insights into neofunctionalization.

To date, only dozens of terpenoid ATs have been reported, approximately twenty of which are involved in triterpenoid biosynthesis, and three have been reported to be involved in limonoid biosynthesis[30,31]. In this study, two acetyltransferases (MaAT8824 and MaAT1704) in chinaberry were screened by WGCNA. Phylogenetic analysis revealed that they occupied a distinct lineage among plant ATs (Supplementary Fig. 30), implying their functional uniqueness. Compared with the reported ATs[32], MaAT8824 had a fairly narrow acceptor spectrum, mainly compounds **1**, **3**, **5**, **6** and **14** with higher affinity towards compound **1**. Since we do not have more analogues available, a broader acceptor spectrum of this enzyme cannot be excluded. On the other hand, the enzyme could only use acetyl-CoA as a donor, which is similar to the known terpenoid acyltransferases[33–36]. The acetyltransferases identified in this study could catalyse limonoid acetylation towards both C12-OH and C3-OH groups. Multiple catalytic sites for acetyltransferases are seldom found among the reported terpenoid acyltransferases[30]. The dual activity of MaAT8824 towards C12-OH and C3-OH groups prompted us to explore its catalytic mechanism by MD simulations, kinetic assays and site-specific mutagenesis. MD simulations, kinetic assays and reaction rate comparisons revealed a preference for C12-OH acetylation. Mutation of W377A resulted in the loss of catalytic activity towards Substrate **1** but not Substrate **6**, indicating the importance of this residue in determining its catalytic activity towards the C12-OH or C3-OH groups of limonoids. Although W377 is reported to be vital for stabilising the tetrahedral transition state through hydrogen bonds in some ATs[28,29,37,38], this mutation of W377A presented slight catalytic activity against compounds **1** and **6** (Fig. 6h), which may be due to the uniqueness of the AT identified in this study since it varied greatly from the reported ATs (Fig. 1). MaAT8824 also presented lower acetylation activity towards compound **5** with both C3 and C12-OH groups. The lower kinetic parameters of compound **5** than those of compound **6** indicate the importance of the acetyl group. By combining the results of the MD simulations, we hypothesised that the absence of this C12-acetyl group results in a greater position shift of compound **5**, which leads to its lower catalytic activity. Moreover, MaAT8824 catalysed compounds **3** and **14** less efficiently than compounds **1** and **5**, respectively, thereby implying that the D-ring in compounds **3** and **14** was probably responsible for this. The crystal structure of MaAT8824 will be critical for clarifying this phenomenon in the future.

In summary, we reported the T2T genome assemblies of neem and chinaberry, revealed the allopatric speciation of the two trees from the uplift of the Himalaya–Hengduan Mountains approximately 18 MYA, and identified key acetyltransferase genes involved in the biosynthesis and divergence of limonoids.

## Methods

### Chemicals

Yeast powder, tryptone, trometamol (Tris), hydrochloride, and agar powder were purchased from Solaibo Technology Co., Ltd. (Beijing, China). All reagents used in this study were at least analytical grade and were purchased from Beihua Fine Chemicals Co., Ltd. (Beijing, China). The reference standards of toosendanin **2**, dihydroniloticin **7**, niloticin **8**, butyrospermol, *S*-allylcysteine and piscidinol A **11** (purity > 98%) were purchased from Yunnan Xili Biotechnology Co., Ltd. (Yunnan, China). The standard isotoosendanin **4** was purchased from Weikeqi Biological Technology (Chengdu, Sichuan, China). 12-Hydroxyamoorastatone **3** was kindly provided by Prof. Qin-Gang Tan (Guilin Medical University). The specific chemical compounds **1**, **5**, **6**, **9** and **14** were synthesised by WuXi AppTec (Tianjin, China).

### Plant materials and sequencing

Fresh plant samples were collected from living *A. indica* and *M. azedarach* plants in Hainan Province, China. The neem trees were approximately 15 to 20 years old, whereas the chinaberry trees were approximately 15 years old. The GPS coordinates of the study site are 19.50437541°N, 109.49729178°E, with an altitude of 124 m. Sample collection occurred in February. Four tissues (stems, leaves, fruits, and branches) from *M. azedarach* and *A. indica* were collected, with three biological replicates for each. High-quality genomic DNA was isolated from fresh leaves using the standard CTAB method[39]. DNA quality and concentration were measured using 0.75% agarose gel electrophoresis and a Qubit 3.0 fluorometer (Life Technologies, Carlsbad, CA, USA).

A Covaris ultrasonic disruptor was used to randomly shear the genomic DNA material. Paired-end libraries with an insert size of approximately 300 bp were prepared, and sequencing was conducted on the DNBSEQ-T7 platform (Beijing Genomics Institute). The SOAPnuke tool (version 2.1.4) (https://github.com/BGI-flexlab/SOAPnuke) was utilised to discard low-quality reads, specifically those containing adaptors, unknown nucleotides, or more than 20% low-quality bases, to obtain clean data for subsequent analyses.

The SQK-ULK001 Ultralong DNA Sequencing Kit was utilised to prepare libraries for Oxford Nanopore ultralong sequencing, resulting in read lengths with an N50 greater than 50 kb. These libraries were subsequently loaded onto primed R9.4 Spot-On Flow Cells. Sequencing was conducted using a PromethION sequencer (Oxford Nanopore Technologies, Oxford, UK) over 48 h runs. Base-calling of the raw data was performed using the Oxford Nanopore GUPPY (version 0.3.0). In addition, high-fidelity (HiFi) data were sequenced on the PacBio Sequel II platform.

High-throughput chromosome conformation capture (Hi-C) sequencing started with chromatin fixation with formaldehyde. In situ capture was performed according to the DNase-based protocol[40]. The HiC libraries of *A. indica* and *M. azedarach* were sequenced in 150 bp paired-end mode on the MGI DNBSEQ-T7 platform (Beijing Genomics Institute).

Total RNA was extracted using the RNAprep Pure Plant Plus Kit (Tiangen Biotech (Beijing) Co., Ltd., China). A cDNA-PCR Sequencing Kit (SQK-PCS109) was used to sequence the cDNA with a PromethION sequencer (Oxford Nanopore Technologies, Oxford, UK). Low-quality data from the raw sequencing reads were filtered out to obtain clean data for subsequent analyses. RSEM[41] was used to estimate the expression level of each gene in terms of fragments per kilobase per million bases (FPKMs).

### Estimation of the genome size

Filtlong (v0.2.0, https://github.com/rrwick/Filtlong) and Porecop (v0.2.2, https://github.com/rrwick/Porechop) were used to filter the original ONT ultralong sequencing data, and PacBio HiFi sequencing data (subreads) were filtered using ccs (option: -min passes 3 -min sn r 2.5 -top passes 60, https://github.com/PacificBiosciences/ccs). The second-generation raw reads, including HiC sequencing reads, were filtered using fastp (v0.21.0)[42]. Based on the BGI short reads of both species, a K-mer analysis ($K = 19$) was performed to estimate the genome size and heterozygosity using GCEpackage (v.1.0.0) and GenomeScope[43].

### De novo assembly and assessment

We used three strategies to obtain the contig-level genome assembly. First, ONT ultralong sequencing data were assembled using NextDenovo[44–47] (v2.5.0, option: read_cutoff = 1 k, blocksize = 1 g, nextgraph_options = -a 1; https://github.com/Nextomics/NextDenovo). Two rounds of error correction were performed on the assembly result using the nanopore sequencing data with Racon[48] (v1.4.11) (https://github.com/isovic/racon), and another two rounds of correction were performed using the DNBSEQ-T7 sequencing data with Pilon[49] (v1.23). Second, the PacBio HiFi sequencing data alone were assembled using Hifiasm[50] (v0.19.8). Third, both the ONT ultralong and PacBio HiFi sequencing data and the Hi-C interaction data were jointly used to obtain two haplotype-resolved assemblies with Hifiasm (v0.19.8). After these three assemblies were evaluated, the best contig-level assembly was chosen for the following analysis, e.g., gap-filling, HiC-assisted scaffolding, and manual correction.

### HiC-assisted pseudochromosome assembly

For pseudochromosome-level scaffolding, the assembly software ALLHiC[51] (v0.9.12) was used for stitching. The final files (.hic and.assembly) were imported into Juicebox[52] (v1.11.08) for visualisation-assisted manual correction. Gaps in the genome were filled using Winnowmap2[53,54] (v1.11, option: $k = 15$, –MD). Finally, the heterozygous sequences of the raw genome assembly were removed using the Purge_haplotigs pipeline (v.1.0.4) to obtain the final assembly result[55].

### Estimation of the rDNA copy number using ddPCR

Droplet digital PCR (ddPCR) was performed to quantify the copy number of rDNA using genomic DNA extracted using QuickExtract DNA Extraction Solution (TSINGKE), and different concentrations of DNA were diluted in nuclease-free $H_2O$ to serve as templates for the ddPCR. ddPCRs were set up using 2×T5 fast qPCR mix, 50×ROX Reference Dye I, primer pairs, and probes. Droplets were generated and analysed using Sniper DQ-24 according to the manufacturer's instructions. The ddPCR cycling conditions were as follows: 60 °C (5 min) and 95 °C (2 min), followed by 42 cycles of 95 °C (15 s) and 60 °C (30 s).

### Repeat and ncRNA annotations

We predicted the LTR sequences and removed redundancy using LTR_FINDER[56] (option: -threads 16 - harvest_out - size 1000000 - time 300, https://github.com/xzhub/LTR_Finder) and LTR_retriever[57] (v2.9.0, option: - threads 16). The genomic repeat sequence was masked using RepeatModeller2[58] and RepeatMask[59] (option: - noLowSimple–$p$-value 0.0001). The genomic tRNA sequences in the genome were identified according to the structural characteristics of tRNAs via tRNAscan-SE[60], and ncRNA sequences in the genome were predicted based on the Rfam database using INFERNAL[61] (v1.1).

### Gene prediction and functional annotation

Gene models for both the *A. indica* and *M. azedarach* genome assemblies were predicted via RNAseq-based transcript mapping, ab initio gene prediction, and homologous gene alignment. The RNAseq data were assembled to obtain unigenes using stringtie2[62] (v2.1.5), and the open reading frames of all the assembled unigenes were predicted with TransDecoder (v5.1.0) (https://github.com/TransDecoder/TransDecoder). Augustus[63] (v3.3.2), Genscan[64] (v1.0), and GlimmerHMM[65] (v3.0.4) were used for the ab initio gene prediction. For homologous gene alignment, the proteins from four related species (*Arabidopsis thaliana, Coptis chinensis, Papaver rhoeas*, and *Aquilegia viridiflora*) were aligned to the assembled *A. indica* and *M. azedarach* genomes using Exonerate[66] (v2.4.0). Finally, MAKER2[67] (v2.31.10) was used to combine the gene sets predicted by these three methods and remove incomplete genes and genes with too short CDSs (CDS length < 150 bp), resulting in the final, nonredundant, and more complete predicted gene set. We employed BUSCO[68] (v.5.2.2) to evaluate the quality of the prediction based on the Embryophyta database. We performed reannotation for genomes of *T. sinensis* and *S. macrophylla* using the same softwares and relative species.

### Phylogenetic analysis and divergence time estimation

All protein sequences of the 19 selected species were aligned using BLASTP[69] (v.2.6.0; parameter: -evalue 1e-5 -outfmt 6), and Orthofinder[70] software was used to perform gene family clustering (v.2.3.12; parameters: -M msa). Single-copy gene families shared by all 19 species were collected to construct the phylogenetic tree. The phylogenetic tree was constructed using RAxML[71] (v.8.2.10) with 100 bootstrap replicates. For the estimation of the divergence times of these plants, MCMCtree[72] was used, with the published divergence times of *Oryza sativa* and *V. vinifera* (125–150 Mya) used to calibrate the divergence times.

### Analysis of whole-genome duplication events

We compared the genomes of *A. indica* and *M. azedarach* with those of 19 other plant species to investigate gene family expansion and contraction. Gene family expansion and contraction analyses were performed using CAFÉ[73] (v.3.1, options: -p 0.05 -t 8 -r 10000 -filter) based on the clustering results.

Whole-genome duplication (WGD) events are typically identified from the Ks (the number of substitutions per synonymous site) age distributions of paralogues. All the amino acid sequences of the paralogues were self-aligned using BLASTP[69] (v.2.6.0, options: -evalue 1e-05 -outfmt 6), and only the best BLASTP results were retained. The syntenic blocks were classified using MCScanX[74] (options: -a -e 1e-5 -s 5 -m 25, https://github.com/wyp1125/MCScanx). Finally, the Ks values were derived from paralogous families using PAML[72].

### Transcriptome analysis and cDNA cloning

For the transcriptome analysis, four tissues (stems, leaves, fruits, and branches) were collected from *M. azedarach* and *A. indica*, with three biological replicates for each. Transcriptome sequencing was performed by Anoroad Gene Technology Co., Ltd. (Beijing, China). Total RNA was extracted from the fruit using a Quick RNA isolation kit (TIANGEN Biotech, Beijing, China) according to the manufacturer's instructions. The full-length cDNAs of *MaAT8824, MaAT1704*, and *AiAT0635* were identified based on the genome and transcriptome sequencing data. The open reading frames were further cloned and inserted into the pET28a vector for the functional analysis. Mutants of MaAT8824 and AiAT0635 were generated by site-specific mutagenesis.

### Weighted gene co-expression network analysis (WGCNA)

The WGCNA package (v1.70-3) was used to conduct the weighted correlation network analysis among 12 transcriptome samples from 4 different tissues for *M. azedarach* and 15 transcriptome samples from 5 different tissues for *A. indica*, namely, fruits, leaves, stems, flowers, and branches, with three biological replicates for each tissue. Cytoscape (v3.9.0) was used to analyse the gene coexpression network and visualisation.

### In vitro assays of the candidate *AT* genes

The recombinant pET28a-AT plasmids were then inserted into *E. coli* BL21 (DE3) for heterologous expression. The strains harbouring the respective plasmids were subsequently grown in LB media supplemented with 100 μg/mL kanamycin (Kan) at 37 °C. Gene expression was induced by the addition of 0.4 mM IPTG when the $OD_{600}$ reached 0.5–0.6, followed by further incubation with shaking at 140 rpm at 16 °C for 20 h. The cell pellets were harvested by centrifugation at 8000 × $g$ for 5 min at 4 °C and then washed with lysis buffer containing 50 mM Tris-HCl and 300 mM NaCl (pH 7.5). The cells were disrupted by

four passes at ~15,000 psi using a high-pressure homogeniser (JNHIO, Juneng Nano & Biotechnology). After cell disruption, the cell debris was removed by centrifugation (12000 × *g* for 20 min at 4 °C), and the resulting supernatant was used as the crude enzyme solution. The reaction mixture (500 µL) used to assess enzymatic activity contained 50 mM Tris-HCl (pH 7.5), 0.02 mM acceptor substrate, 0.1 mM acetyl-CoA, and 450 µL of the crude protein solution. Each reaction mixture was incubated at 30 °C for 4 h, and the reaction was terminated by the addition of 500 µL of methanol. The extracts were evaporated and reconstituted in 100 µL of methanol and analysed by LC–qTOF–MS/MS. The primers were detailed in Supplementary Table 10.

### Transient expression in *N. benthamiana*

The recombinant pEAQ-HT-AT plasmids were constructed and subsequently electroporated into *Agrobacterium tumefaciens* GV3101. The strains harbouring the respective plasmids were subsequently grown in LB media at 28 °C until the $OD_{600}$ reached 1.0. The cell pellets were harvested by centrifugation at 8000 × *g* for 10 min and then resuspended in MMA solution (10 mM MES, 10 mM $MgCl_2$, and 150 µM acetosyringone, pH 5.6) to an $OD_{600}$ of 0.2. The mixture was subsequently incubated in the dark at 28 °C for an hour and infiltrated into the leaves of 4-week-old *N. benthamiana* plants. The 0.02 mM substrate was infiltrated after 3 days. The infected leaves of the same weight and state were collected, freeze-dried, and crushed into a powder weighing 10 mg. The infected leaves were extracted with methanol, sonicated for 1 h and then centrifuged at 8000 × *g* for 30 min. The extracts were evaporated, reconstituted in 100 µL of methanol and analysed using LC–qTOF–MS/MS.

### Purification and kinetic assay of MaAT8824

The His-tagged MaAT8824 was purified from the soluble lysate using Ni-NTA affinity chromatography (His-Bind Resin, Novagen). Using an empty gravity flow column (Bio-Rad), the filtered crude enzyme protein was incubated with His-Bind resin for 2 h at 4 °C. The resin was washed twice with wash buffer (50 mM Tris-HCl and 300 mM NaCl, pH 7.5). The His-tagged proteins were subsequently eluted from the resin using elution buffer (50 mM Tris-HCl, 300 mM NaCl, and 20 mM imidazole, pH 7.5). For the kinetic assay of MaAT8824 using compounds **1**, **5**, and **6** as substrates, the reaction conditions were tested using a 500 µL reaction mixture containing 50 mM Tris-HCl (pH 7.5), 300 mM NaCl, 0.01-1 mM substrates, 1.0 mM acetyl-CoA, and 0.1 µg/mL purified proteins at 37 °C for 30 min. Methanol (1 mL) was added to the mixture for the LC-QQQ-MS analysis.

### *S*-allylcysteine and limonoid extraction and detection

A total 2 mg samples of stems and leaves was weighed, and extracted with 40 mL of 20% ethanol to detect the *S*-allylcysteine from *M. azedarach*, *A. indica, T. sinensis*, and *C. sinensis*. A total of 2 mg samples of fruits, stems, leaves, and branches was weighed and extracted with 20 mL of 95% ethanol to detect the limonoids in all tissues from *M. azedarach* and *A. indica*. The mixed sample was sonicated for 1 h and then centrifuged at 8000 × *g* for 30 min. The supernatant was collected after each centrifugation step, and the extraction, sonication, and centrifugation cycles were repeated three times. The supernatant was subsequently collected, evaporated to dryness, and then redissolved in 3 mL of methanol. The organic phase was filtered through a 0.22 µm membrane filter and concentrated again before analysis using LC-QQQ-MS.

### MD analysis of MaAT8824

The structures of MaAT8824 and AiAT0635 were modelled using the Alphafold2 server[75]. Multiple sequence alignment (MSA) searches were conducted with MMseqs2, and templates for the structural prediction were sourced from the PDB70 database. The ligand was obtained from PubChem. Docking studies were conducted using AutoDock 4.0, which employs a semiflexible docking method to position the ligand within the active site of the rigid protein. Fifty docking conformations were generated, from which the structure exhibiting the lowest docking energy and an appropriate catalytic distance near the active site was selected as the initial molecular dynamics (MD) structure[76]. The hydroxyl group on the substrate used in the reaction was set to a deprotonated state, whereas H164 was configured to a diprotonated state to align with the structure of the reaction intermediate.

Molecular dynamics (MD) simulations were performed using the GROMACS programme[77,78], with the protein modelled using the CHARMM36 force field. The integration step was set to 2 fs. Following minimisation, the temperature of each system was gradually increased to 300 K in the NVT ensemble over 200 ps, utilising the V-rescale thermostat for temperature control[79]. Subsequently, MD simulations were conducted in the NPT ensemble for 400 ps ($T = 300$ K and $P = 1$ atm), with pressure regulated by the Parrinello–Rahman barostat[80]. Finally, a molecular dynamics simulation was executed for 10 ns, and the system coordinates were recorded every 10 ps. Each system was independently simulated three times. During the MD simulation, the substrate remained in the reaction intermediate state to accurately characterise the reaction distance. GROMACS 5.1.5 and VMD 1.9.2[81] were employed to analyse the MD simulation trajectories, whereas PyMOL 2.5.4 was used for the active site analysis.

### Metabolite profiling using LC-qTOF-MS and LC-QQQ-MS

The LC-qTOF-MS analysis was conducted using an Agilent 1290 Infinity II UHPLC system coupled with a 6546 qTOF (Agilent). For MS acquisition, the data scan range of the mass–charge ratio (*m/z*) was set from 100 to 1500. Chromatographic separation was performed on an Acquity UPLC BEH C18 column (1.7 µm, 2.1 × 100 mm, Waters) at a column temperature of 45 °C and a flow rate of 0.3 mL/min. The mobile phases consisted of 0.1% formic acid in water (A) and acetonitrile (B). The gradient conditions were as follows: 0–3.0 min, 30% to 30.5% B; 3.0–4.0 min, 30.5% to 30.5% B; 4.0–15.0 min, 30.5% to 35% B; 15.0–19.0 min, 35% to 100% B; 19.0–22.0 min, 100% B; and 22.0–23.0 min, 100% to 30% B.

The LC-QQQ-MS analysis was performed on an Agilent UPLC-QQQ6470 (Agilent, USA) in multiple reaction monitoring (MRM) modes (Supplementary Table 11). Chromatographic separation was achieved using an Acquity UPLC BEH C18 column (1.7 µm, 2.1 × 100 mm, Waters), with a sample volume of 5 µL loaded onto the column. The mobile phase consisted of solution A (0.1% formic acid in water) and solution B (acetonitrile), with a flow rate of 0.3 mL/min. The column temperature was maintained at 35 °C. Data acquisition was performed using the Agilent UPLC-QQQ6470 in MRM mode. The gradient conditions for the mobile phase were as follows: 0–3.0 min, 25% to 30% B; 3.0–10.0 min, 30% to 40% B; 10.0–16.0 min, 40% to 95% B; and 16.0–20.0 min, 95% B. MS acquisition was performed using an ESI ion source operating in positive ion mode, and the data were processed with MassHunter 10.0 software. Metabolites were all quantified with reference to the corresponding standard curves.

### Reporting summary

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

## Data availability

The raw data of genome and transcriptome in this study have been deposited in the Genome Sequence Archive in National Genomics Data Centre, China National Centre for Bioinformation under accession CRA019071. The genome assembly and gene annotation have been deposited in Genome Warehouse in National Genomics Data Centre, China National Centre for Bioinformation under accession PRJCA030181. Source data are provided in this paper.

## References


1. Tan, Q. G. & Luo, X. D. Meliaceous limonoids: Chemistry and biological activities. *Chem. Rev.* **111**, 7437–7522 (2011).
2. Luo, J., Sun, Y., Li, Q. & Kong, L. Research progress of meliaceous limonoids from 2011 to 2021. *Nat. Prod. Rep.* **39**, 1325–1365 (2022).
3. Krishnan, N. M. et al. A draft of the genome and four transcriptomes of a medicinal and pesticidal angiosperm *Azadirachta indica*. *BMC Genomics* **13**, 1–13 (2012).
4. Krishnan, N. M., Jain, P., Gupta, S., Hariharan, A. K. & Panda, B. An improved genome assembly of *Azadirachta indica* A. Juss.*G3* **6**, 1835–1840 (2016).
5. Cui, G. et al. Meliaceae genomes provide insights into wood development and limonoids biosynthesis. *Plant Biotechnol. J.* **21**, 574–590 (2023).
6. Ji, Y. T. et al. Long read sequencing of *Toona sinensis* (A. Juss) Roem: A chromosome-level reference genome for the family Meliaceae. *Mol. Ecol. Resour.* **21**, 1243–1255 (2021).
7. Hodgson, H. et al. Identification of key enzymes responsible for protolimonoid biosynthesis in plants: Opening the door to azadirachtin production. *Proc. Natl. Acad. Sci. USA* **116**, 17096–17104 (2019).
8. De La Peña, R. et al. Complex scaffold remodeling in plant triterpene biosynthesis. *Science* **379**, 361–368 (2023).
9. National Research Council (US) Panel on the Applications of Biotechnology to Traditional Fermented Foods. *Applications of Biotechnology to Fermented Foods: Report of an Ad Hoc Panel of the Board on Science and Technology for International Development*. (National Academies Press (US), Washington (DC), 1992).
10. Livingstone, K. & Rieseberg, L. Chromosomal evolution and speciation: a recombination- based approach. *New Phytol.***161**, 107–112 (2004).
11. Kirkpatrick & Genetics, M. J. Chromosome inversions, local adaptation and speciation. *Genetics* **173**, 419–434 (2006).
12. Stevison, L. S., Hoehn, K. B. & Noor, M. A. F. Evolution Effects of inversions on within- and between-species recombination and divergence. *Genome Biol. Evol.* **3**, 830–841 (2011).
13. Wellenreuther, M. & Bernatchez, L. Eco-evolutionary genomics of chromosomal inversions. *Trends Ecol. Evol.* **33**, 427–440 (2018).
14. Tuskan, G. A. et al. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* **313**, 1596–1604 (2006).
15. Tapponnier, P. et al. Oblique stepwise rise and growth of the Tibet Plateau. *Science* **294**, 1671–1677 (2001).
16. Royden, L. H., Burchfiel, B. C. & van der Hilst, R. D. The geological evolution of the Tibetan Plateau. *Science* **321**, 1054–1058 (2008).
17. Hall, R. & Morley, C. K. *Sundaland Basins* (2004).
18. Daly, D., Harley, M., Martínez-Habibe, M. & Weeks, A. *Flowering plants. Eudicots: Sapindales, Cucurbitales, Myrtaceae*. (Springer Berlin-Heidelberg, Berlin, 2011).
19. Lancaster, J. E. & Shaw, M. L. γ-Glutamyl peptides in the biosynthesis of *S*-alk (en) yl-L-cysteine sulphoxides (flavour precursors) in Allium. *Phytochemistry* **28**, 455–460 (1989).
20. Sun, X. et al. A chromosome-level genome assembly of garlic (*Allium sativum*) provides insights into genome evolution and allicin biosynthesis. *Mol. Plant* **13**, 1328–1339 (2020).
21. Carpinella, M. C., Defago, M. T., Valladares, G. & Palacios, S. M. Antifeedant and insecticide properties of a limonoid from *Melia azedarach* (Meliaceae) with potential use for pest management. *J. Agric. Food Chem.* **51**, 369–374 (2003).
22. Zhou, J. B., Minami, Y., Yagi, F., Tadera, K. & Naktani, M. J. Antifeeding limonoids from *Melia toosendan*. *Heterocycles* **9**, 1781–1786 (1997).
23. Nakatani, M. et al. Nimbolinins, C-Seco Limonoids from the Fruits of *Malia toosendan*. *Heterocycles* **53**, 689–696 (2000).
24. Bhambhani, S. et al. Transcriptome and metabolite analyses in *Azadirachta indica*: identification of genes involved in biosynthesis of bioactive triterpenoids. *Sci. Rep.* **7**, 5043 (2017).
25. Hansen, D. J., Cuomo, J., Khan, M., Gallagher, R. T. & Ellenberger, W. P. *Natural and Engineered Pest Management Agents*. (ed. Society, A. C.) (ACS Symposium Series, Washington, DC; 1994).
26. Wang, X., Wang, C. & Wang, Z. Determination of toosendanin in rat plasma by ultra-performance liquid chromatography-electrospray ionization-mass spectrometry and its application in a pharmacokinetic study. *Biomed. Chromatogr.* **27**, 222–227 (2013).
27. Birchler, J. A. & Hua, Y. J. The multiple fates of gene duplications: Deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. *Plant Cell* **34**, 2466–2474 (2022).
28. Walker, A. M. et al. Elucidation of the structure and reaction mechanism of sorghum hydroxycinnamoyltransferase and its structural relationship to other coenzyme A-dependent transferases and synthases. *Plant Physiol.* **162**, 640–651 (2013).
29. Wang, L. et al. Characterization and structure-based protein engineering of a regiospecific saponin acetyltransferase from Astragalus membranaceus. *Nat. Commun.* **14**, 5969 (2023).
30. Wang, L., Chen, K., Zhang, M., Ye, M. & Qiao, X. Catalytic function, mechanism, and application of plant acyltransferases. *Crit. Rev. Biotechnol.* **42**, 125–144 (2021).
31. Zhang, F. et al. Herbal multiomics provide insights into gene discovery and bioproduction of triterpenoids by engineered microbes. *J. Agric. Food Chem.* **73**, 47–65 (2025).
32. Walker, K., Fujisaki, S., Long, R. & Croteau, R. Molecular cloning and heterologous expression of the C-13 phenylpropanoid side chain-CoA acyltransferase that functions in Taxol biosynthesis. *Proc. Natl. Acad. Sci. USA* **99**, 12715–12720 (2002).
33. Walker, K., Long, R. & Croteau, R. The final acylation step in Taxol biosynthesis: Cloning of the taxoid C13-side-chain *N*-benzoyltransferase from Taxus. *Proc. Natl. Acad. Sci. USA* **99**, 9166–9171 (2002).
34. Walker, K. D. & Croteau, R. B. Molecular Cloning of a 10-Deacetylbaccatin III-10-0-acetyl Transferase cDNA from Taxus and Functional Expression in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **97**, 583–587 (2000).
35. Choi, H. S., Han, J. Y., Cheong, E. J. & Choi, Y. E. Characterization of a pentacyclic triterpene acetyltransferase involved in the biosynthesis of taraxasterol and ψ-taraxasterol acetates in lettuce. *Front. Plant Sci.* **12**, https://doi.org/10.3389/fpls.2021.788356 (2022).
36. Shang, Y. et al. Biosynthesis, regulation, and domestication of bitterness in cucumber. *Science* **346**, 1084 (2014).
37. Unno, H. et al. Structural and mutational studies of anthocyanin malonyltransferases establish the features of BAHD enzyme catalysis. *J. Biol. Chem.* **282**, 15812–15822 (2007).
38. Ma, X., Koepke, J., Panjikar, S., Fritzsch, G. & Stockigt, J. Crystal structure of vinorine synthase, the first representative of the BAHD superfamily. *J. Biol. Chem.* **280**, 13576–13583 (2005).
39. Porebski, S., Bailey, L. G. & Baum, B. R. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Mol. Biol. Rep.* **15**, 8–15 (1997).
40. Ramani, V. et al. Mapping 3D genome architecture through in situ DNase Hi-C. *Nat. Protoc.* **11**, 2104–2121 (2016).
41. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
42. Chen, S. F., Zhou, Y. Q., Chen, Y. R. & Gu, J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, 884–890 (2018).
43. Vurture, G. W. et al. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
44. Sun, J., Li, R., Chen, C., Sigwart, J. D. & Kocot, K. M. Benchmarking Oxford Nanopore read assemblers for high-quality

molluscan genomes. *Philos. Trans. R. Soc. B Biol. Sci.* **376**, 20200160 (2021).
45. Guiglielmoni, N., Houtain, A., Derzelle, A., Van Doninck, K. & Flot, J. F. Overcoming uncollapsed haplotypes in long-read assemblies of non-model organisms. *BMC Bioinformatics* **22**, 303 (2021).
46. Zhang, T. et al. Comparison of long-read methods for sequencing and assembly of lepidopteran pest genomes. *Int. J. Mol. Sci.* **24**, 649 (2022).
47. Mochizuki, T. et al. A practical assembly guideline for genomes with various levels of heterozygosity. *Brief. Bioinform.* **24**, 1–13 (2023).
48. Vaser, R., Sovic, I., Nagarajan, N. & Sikic, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
49. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *Plos ONE* **9**, e112963 (2014).
50. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
51. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).
52. Robinson, J. T. et al. Juicebox.js provides a cloud-based visualization system for Hi-C data. *Cell Syst.* **6**, 256–258.e251 (2018).
53. Jain, C., Rhie, A., Hansen, N. F., Koren, S. & Phillippy, A. M. Long-read mapping to repetitive reference sequences using Winnowmap2. *Nat. Methods* **19**, 705–710 (2022).
54. Jain, C. et al. Weighted minimizer sampling improves long read mapping. *Bioinformatics* **36**, 111–118 (2020).
55. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460 (2018).
56. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
57. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol* **176**, 1410–1422 (2018).
58. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
59. Tarailo-Graovac, M. & Chen, N. *Current Protocol in Bioinformatics* (2009).
60. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
61. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
62. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
63. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–W467 (2005).
64. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
65. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
66. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
67. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
68. Waterhouse, R. M. et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).
69. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
70. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
71. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
72. Yang, Z. H. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
73. De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269–1271 (2006).
74. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
75. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
76. Cui, Z. et al. Green biomanufacturing promoted by automatic retrobiosynthesis planning and computational enzyme design. *Chin. J. Chem.Eng.* **41**, 6–21 (2022).
77. Abraham, M. J. et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1**, 19–25 (2015).
78. Zhang, S. et al. Enzyme self-aggregation in supramolecular self-assembly of glucose oxidase and catalase: Insight from molecular dynamics simulation based on coarse-grained method. *Chem. Phys.* **552**, 111366 (2022).
79. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, https://doi.org/10.1063/1.2408420 (2007).
80. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).
81. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).
82. Du, Y. et al. Genomic analysis based on chromosome-level genome assembly reveals an expansion of terpene biosynthesis of Azadirachta indica. *Front. Plant Sci.* **13**, 853861 (2022).
83. Kuravadi, N. A. et al. Comprehensive analyses of genomes, transcriptomes and metabolites of neem tree. *PeerJ* **3**, e1066 (2015).

## Acknowledgements

This work was supported by the Scientific and Technological Innovation Project of China Academy of Chinese Medical Sciences (CI2024C006YN, CI2023E002, CI2021A04111) (C.W.), the Fundamental Research Funds for the Central Public Welfare Research Institutes (ZZ13-YQ-040) (C.W.), and the National Natural Science Foundation of China 32372305 (C.W.), 32300327(J.L.).

## Author contributions

Caixia Wang, Yujun Zhang, and Shilin Chen conceived and designed the research. Jia Liu, Xinyao Su and Zhennan Wang performed the experiments. Zhennan Wang, Caixia Wang, Yujun Zhang and Liang Leng analysed the results. Jiarou Liu and Feng Zhang collected the materials and performed the experiments. Caixia Wang wrote the manuscript. Yujun Zhang reviewed and edited the manuscript. All authors have read and approved the manuscript for publication.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-57722-9.

**Correspondence** and requests for materials should be addressed to Shilin Chen, Yujun Zhang or Caixia Wang.

**Peer review information** *Nature Communications* thanks Zhihua Liao, Sumit Ghosh and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.