the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2022) **109,** 891–908 doi: 10.1111/tpj.15600

Chromosome-level assembly of the *Neolamarckia cadamba* genome provides insights into the evolution of cadambine biosynthesis

Xiaolan Zhao[1,2,†], Xiaodi Hu[3,†], Kunxi OuYang[1,2,†], Jing Yang[1,2,4,†], Qingmin Que[1,2], Jianmei Long[1,2], Jianxia Zhang[1,2], Tong Zhang[1,2], Xue Wang[1,2], Jiayu Gao[1,2], Xinquan Hu[1,2], Shuqi Yang[1,2], Lisu Zhang[1,2], Shufen Li[5], Wujun Gao[5], Benping Li[3], Wenkai Jiang[3], Erik Nielsen[1,2,6], Xiaoyang Chen[1,2,*] and Changcao Peng[1,2,*]

[1]*State Key Laboratory for Conservation and Utilization of Subtropical Agro-bioresources, South China Agricultural University, Guangzhou 510642, China,*

[2]*Guangdong Key Laboratory for Innovative Development and Utilization of Forest Plant Germplasm, College of Forestry and Landscape Architecture, South China Agricultural University, Guangzhou 510642, China,*

[3]*Novogene Bioinformatics Institute, Building 301, Zone A10 Jiuxianqiao North 13 Road, Chaoyang District, Beijing 100083, China,*

[4]*School of Chinese Medicinal Resource, Guangdong Pharmaceutical University, Guangzhou 510006, China,*

[5]*College of Life Sciences, Henan Normal University, Xinxiang 453007, China, and*

[6]*Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, Michigan 48109, USA*

Received 7 July 2021; revised 1 November 2021; accepted 18 November 2021; published online 22 November 2021.
*For correspondence (e-mails ccpeng@scau.edu.cn; xychen@scau.edu.cn).
†These authors contributed equally to this work.

SUMMARY

Neolamarckia cadamba **(Roxb.), a close relative of** ***Coffea canephora*** **and** ***Ophiorrhiza pumila*****, is an important traditional medicine in Southeast Asia. Three major glycosidic monoterpenoid indole alkaloids (MIAs), cadambine and its derivatives 3β-isodihydrocadambine and 3β-dihydrocadambine, accumulate in the bark and leaves, and exhibit antimalarial, antiproliferative, antioxidant, anticancer and anti-inflammatory activities. Here, we report a chromosome-scale** ***N. cadamba*** **genome, with 744.5 Mb assembled into 22 pseudochromosomes with contig N50 and scaffold N50 of 824.14 Kb and 29.20 Mb, respectively. Comparative genomic analysis of** ***N. cadamba*** **with** ***Co. canephora*** **revealed that** ***N. cadamba*** **underwent a relatively recent whole-genome duplication (WGD) event after diverging from** ***Co. canephora*****, which contributed to the evolution of the MIA biosynthetic pathway. We determined the key intermediates of the cadambine biosynthetic pathway and further showed that NcSTR1 catalyzed the synthesis of strictosidine in** ***N. cadamba*****. A new component, epoxystrictosidine (C27H34N2O10,** ***m/z*** **547.2285), was identified in the cadambine biosynthetic pathway. Combining genome-wide association study (GWAS), population analysis, multi-omics analysis and metabolic gene cluster prediction, this study will shed light on the evolution of MIA biosynthetic pathway genes. This** ***N. cadamba*** **reference sequence will accelerate the understanding of the evolutionary history of specific metabolic pathways and facilitate the development of tools for enhancing bioactive productivity by metabolic engineering in microbes or by molecular breeding in plants.**

Keywords: genome, cadambine biosynthesis, strictosidine synthase, ***Neolamarckia cadamba*****, medicinal plant, evolution.**

INTRODUCTION

The evergreen tropical tree *Neolamarckia cadamba* (Roxb.) Bosser (Rubiaceae), commonly known as Kadamba or Kodom, belongs to the Rubiaceae family, which is the fourth largest family of angiosperms, consisting of more than 660 genera and 11 000 species (Razafimandimbison, 2002, Robbrecht and Manen, 2006). The Rubiaceae family is also noted for the production of important plant alkaloids, which includes well-known plant species such as *Coffea canephora* and *Ophiorrhiza pumila* (Kai et al., 2015; De Luca et al., 2014; Rai et al., 2021; Sadre et al., 2016; Tran et al., 2018). In 1972 *N. cadamba* was called a 'miracle tree' by the World Forestry Congress (WFC) for its considerable economic potential as a fast-growing timber wood and

traditional medicinal resource in tropical and subtropical regions (Dwevedi et al., 2014; Pandey and Negi, 2016). The stems, bark and leaves of *N. cadamba* have been widely used to treat a number of diseases, such as diabetes, anemia, stomatitis, leprosy, cancer and a variety of infectious diseases in Southeast Asia (Pandey and Negi, 2016). Although in-depth studies to clarify the active metabolites responsible for the various pharmacological activities attributed to *N. cadamba* are lacking, recent studies have pinpointed that its three major glycosidic monoterpenoid indole alkaloids (MIAs), cadambine and its derivatives 3β-isodihydrocadambine and 3β-dihydrocadambine, exhibit antimalarial, antiproliferative, antioxidant, anticancer and anti-inflammatory activities (Chandel et al., 2014, 2017; Dwevedi et al., 2014; Yuan et al., 2020).

Since its original discovery in *N. cadamba* (syn. *Anthocephalus chinensis*), cadambine was also found to accumulate in *Emmenopterys henryi*, *Haldina cordifolia* and *Uncaria* species in the Rubiaceae family (Chen et al., 2020; Handa et al., 1983; Wang et al., 2019; Wu et al., 2013). The structure of cadambine was first determined by Handa and co-workers in 1983 and its biosynthesis was deduced to be derived from strictosidine, in which C-18 is cyclized to N-4, with an ether bridge linking C-3 and C-19 (Handa et al., 1983). Most MIAs originate from the common precursor 3-α(*S*)-strictosidine formed by stereospecific condensation of the indole metabolite tryptamine and the end product secologanin in the iridoid (also called secoiridoid) branch (De Luca et al., 2014). However, it remains unknown whether this seco-iridoid pathway exists in *N. cadamba*. Therefore, determining the key intermediate strictosidine and characterizing the functional strictosidine synthase (STR) is required to elucidate the cadambine biosynthetic pathway in *N. cadamba*.

Here we report a chromosome-level genome assembly of *N. cadamba* ($2n = 44$ chromosomes) obtained through a combination of Illumina and PacBio data platforms. We used a high-throughput chromosome conformation capture (Hi-C) map (Burton et al., 2013) to cluster the majority of the assembled contigs onto 22 pseudochromosomes. The *N. cadamba* genome was compared with *Co. canephora* and 12 other available plant genomes to investigate whole-genome duplication (WGD) events and the expansion/contraction of gene families. We further determined the key intermediate strictosidine, the known MIAs 3α-dihydrocadambine and cadambine and a new component, epoxystrictosidine, by mass spectrometry and NMR spectra, and characterized the first 'Pictet-Spenglerase' NcSTR1 in *N. cadamba*. A total of 112 *N. cadamba* accessions collected from Southeast Asia were sequenced to discover more loci and candidate genes for cadambine biosynthesis based on genome-wide association study (GWAS) and population-level analysis. This study revealed the evolution of cadambine biosynthesis in *N. cadamba* and is likely to provide additional insight into plant specialized metabolites.

RESULTS

Genome sequencing, assembly and annotation

The genome size of *N. cadamba* ($2n = 2x = 44$ chromosomes) was estimated to be approximately 754 Mb, with 0.69% heterozygosity and 54.29% repetition, based on the *k*-mer distribution analysis (Figure S1; Table S1). We obtained a total of 92.97 Gb subreads (123.25 ×) generated from the PacBio Sequel platform, plus another 52.30 Gb reads (69.33 ×) from the Illumina platform, 184.45 Gb 10X Genomics data (244.52 ×) and 89.59Gb Hi-C data (118.81 ×) (Table S2). FALCON (Chin et al., 2016) was used for the initial assembly of the PacBio reads, which were then polished and error-corrected with both PacBio and Illumina reads. 10X Genomics data was used to anchor contigs into scaffolds. Finally, using LACHESIS, the assembled scaffolds were anchored to 22 pseudochromosomes based on Hi-C data (Figure 1a). A high-quality chromosome-level genome assembly of *N. cadamba* was obtained with a total length of 744.5 Mb, a contig N50 of 824.14 Kb and a scaffold N50 of 29.20 Mb (Figure 1a; Table 1). The total length of the assembly was 744.5 Mb, which represents 98.7% of the estimated genome size.

To evaluate the quality of the assembly, we first mapped the Illumina reads back to the scaffolds, with a mapping rate of 98.11% and a coverage rate of 94.43%, respectively (Table S3). Second, we also evaluated the assembly using 1614 Benchmarking Universal Single Copy Orthologs (BUSCO) genes from embryophyta (Simao et al., 2015) and 248 highly conserved core eukaryotic genes (CEGs) (Parra et al., 2007), which showed that 1563 genes (96.8%) were annotated and 243 genes (98.0%) were identified in our assembly, respectively (Table S4). Third, the reads of RNA-seq data from 24 samples were mapped to the genome assembly using Hisat2 (Kim et al., 2015a, 2019). The alignment rate of 23 of these RNA samples was over 95%, with the remaining sample aligning at 89.18% (Table S5). Taken together, these results indicate that the assembly of *N. cadamba* has high accuracy and completeness.

Using a combination of ab initio and evidence-based methods, we predicted a total of 35 461 protein-coding genes with an average gene length of 3489 base pairs and an average of 4.7 exons per gene (Table 1). Approximately 96.4% of the genes with shared homology with known genes in NR, Swiss-Prot, KEGG and InterPro databases were functionally annotated (Table S6). In addition, we performed homology searches and annotated non-coding RNA (ncRNA) genes (Table S7), yielding 666 transfer RNA (tRNA) genes, 1642 ribosomal RNA (rRNA, *5S*, *5.8S*, *18S*, *28S*) genes, 2701 small nuclear RNA (snRNA) genes and 1053 microRNA (miRNA) genes.

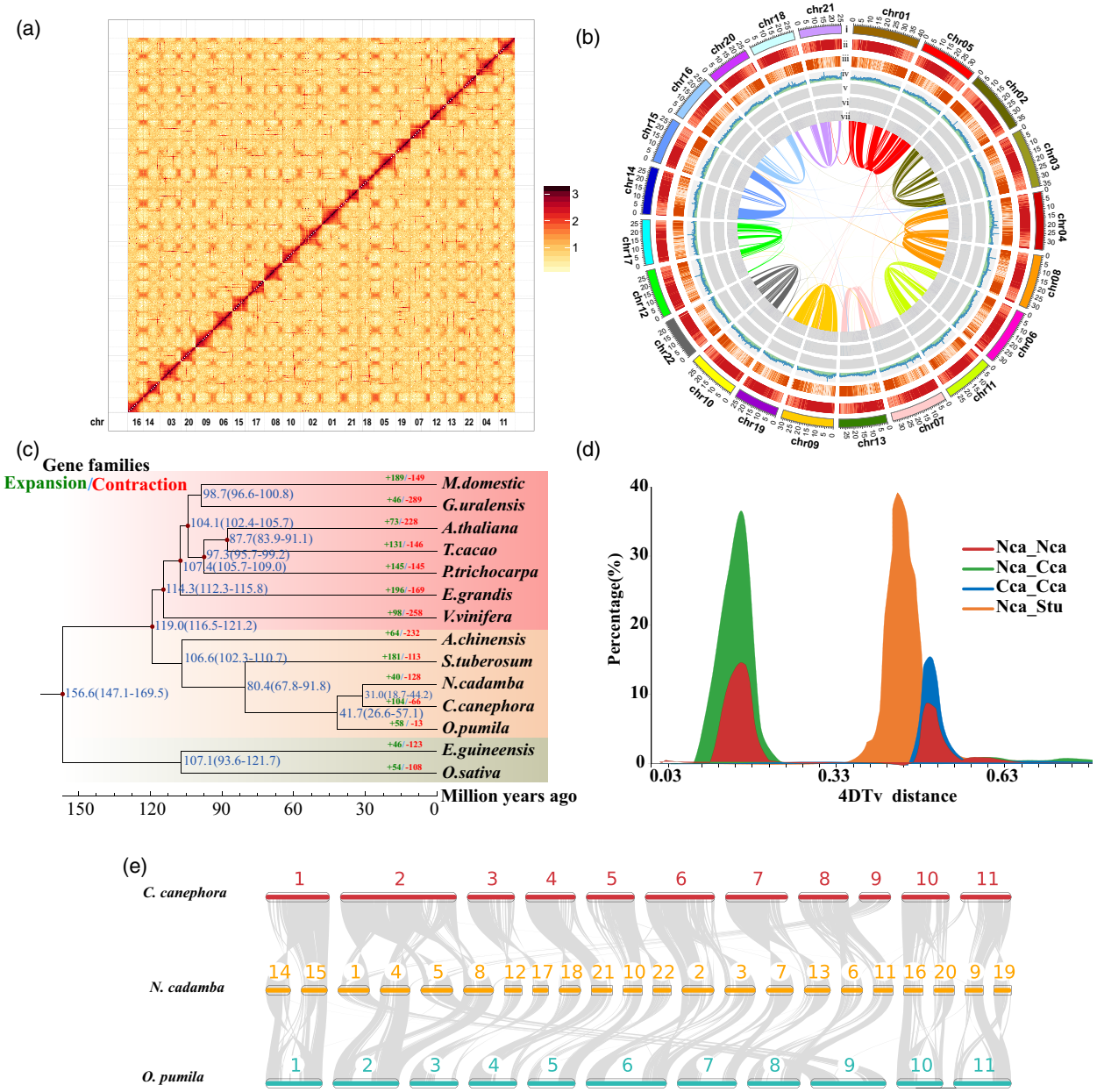

Figure 1. High-throughput chromosome conformation capture (Hi-C) map and overview of the genomic features of the 22 *Neolamarckia cadamba* pseudochromosomes and evolutionary analyses. (a) Hi-C map of the *N. cadamba* genome showing genome-wide all-by-all interactions. (b) Characteristics of the 22 chromosomes of *N. cadamba*. From outermost to innermost layers: (i) circular representation of the 22 chromosomes; (ii–vi) densities of transposable element, gene, GC, miRNA and tRNA; and (vii) densities of snRNA density and syntenic blocks (the densities were calculated in 100-Kb windows.) (c) Phylogenetic tree with 402 single-copy orthologs from 14 species identified by OrthoMCL to show divergence times and expanded/contracted gene families. (d) Distribution of 4DTv distance of homologous genes from *N. cadamba* (Nca), *Coffea canephora* (Cca) and *Solanum tuberosum* (Stu). (e) Synteny blocks between *N. cadamba*, *Co. canephora* and *Ophiorrhiza pumila*.

Repeat annotation revealed that 52.9% (394.1 Mb) of the assembled *N. cadamba* genome comprises transposable elements (TEs) (Figure 1b; Table S8). Retrotransposons were found to be the dominant class of repeat elements (48.2%), whereas DNA transposons account for 2.89% of the genome.

Evolution of the *N. cadamba* genome and comparative genomic analysis

To classify gene families in the *N. cadamba* genome, ORTHOMCL (Li et al., 2003) was used to infer proteins from all 14 plant species (Table S9), generating a total of 32 185 orthologous gene families and 402 single-copy orthologous

Table 1 Global statistics for *Neolamarckia cadamba* genome assembly and annotation

	Number	Size
Genome assembly		
Total contigs	2881	741.90 Mb
Contig N50	225	824.14 Kb
Contig N90	1023	136.8 Kb
Total scaffolds	807	744.45 Mb
Scaffold N50	11	29.20 Mb
Scaffold N90	22	24.51 Mb
Pseudochromosomes	22	744.5 Mb
Genome annotation		
Predicted protein-coding genes	35 461	
Average gene length (bp)		3489.6 bp
Average CDS length (bp)		1151.7 bp
Average exons per gene	4.7	
Average exon length (bp)		245.1 bp
Average intron length (bp)		632.0 bp

genes shared across 14 species. Phylogenetic analysis revealed that *N. cadamba* diverged from *Solanum tuberosum* (Solanaceae), *O. pumila* (Rubiaceae) and *Co. canephora* (Rubiaceae) at around 80.4, 41.7 and 31.0 mya, respectively (Figure 1c). Compared with *Co. canephora*, 40 gene families underwent expansion in *N. cadamba* (Figure 1c). Functional annotation of these expanded genes demonstrated that 161 Gene Ontology (GO) terms and 10 KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways were significantly enriched (false-discovery rate, FDR, cut-off of <0.05) and were involved in defense and sugar metabolism. Among defense response functions, there is a clear expansion of G protein binding site resistance and defense response genes in the *N. cadamba* genome. Another interesting expanded gene family is the GH1 β-glucosidase (BGLU) family, members of which are not only involved in starch and sucrose metabolism, but also play a role in the biosynthesis of phenylpropanoid and secondary metabolites (Ketudat Cairns & Esen, 2010, Xia et al., 2012). These expanded gene families may reflect the rapid growth, synchronized metabolite synthesis and specific adaptations to tropical environments for *N. cadamba* (Figure 1c).

Synteny analysis revealed that two peaks (4DTv distances of approx. 0.17 and 0.51) were observed in the *N. cadamba* genome (Figure 1d). All gene pairs showed a shallow peak at 0.51, likely reflecting a gamma triplication event (whole-genome triplication, WGT-γ) that occurred approximately 70 mya in core eudicots (Paterson et al., 2004). The 4DTv distribution also recovered the WGT-γ in *Co. canephora*, consistent with the previous findings (Denoeud et al., 2014; Hu et al., 2019). Another peak at 0.17 indicated that *N. cadamba* underwent a relatively recent WGD event after diverging from *Co. canephora* (Figure 1d), which was further supported by the distribution of the synonymous substitution rate (K_s) (Figure S2). Intergenomic colinearity analysis demonstrated a 2:1 syntenic relationship between *N. cadamba* and *Co. canephora*, and 186 syntenic blocks were identified in *N. cadamba* by comparison with the *Co. canephora* genome (Figure 1e). KEGG pathway analysis revealed that the duplicated genes from the recent WGD were enriched with terms such as 'glucose metabolism' and 'terpene synthase' (Table S10).

In this study, we also conducted a positive selection analysis using the genomic sequences of *N. cadamba* and three close relatives. A total of 443 genes were found to possibly be under positive selection (PSGs, $P < 0.01$, FDR < 0.05) using the branch-site model of PAML (Yang, 2007). KEGG functional classification of the 443 PSGs (Table S11) showed that the associated categories included '*N*-Glycan biosynthesis', 'Glycolysis/Gluconeogenesis', 'Starch and sucrose metabolism' and 'Plant–pathogen interaction'.

Characterization of cadambine and the key intermediates

We first confirmed the structure of cadambine by quadrupole time-of-flight (Q-TOF) liquid chromatography with tandem mass spectrometry (LC-MS/MS) (Figures 2a and S3) (Chandel et al., 2012, 2017) and NMR spectra (Figure S3) (Handa et al., 1983). The standard tryptamine, 3α-dihydrocadambine (Takayama et al., 2003) and epoxystrictosidine were also analyzed by high-resolution mass spectrometry(Figures 2a and S3). Second, the extracts from the leaves and bark of *N. cadamba* were analyzed, and this analysis revealed that *N. cadamba* accumulated tryptamine (Figure 2c,f) in the leaves, and strictosidine (Figure 2c,g), epoxystrictosidine (Figure 2d,i), 3α-dihydrocadambine (Figure 2d,h) and cadambine (Figure 2e,j) in the bark. Two compounds were resolved with the same *m*/*z* values (547.2285) (Figure 2h,i): one was matched with the peak at 12.30 min, as standard for 3α-dihydrocadambine ($C_{27}H_{34}N_2O_{10}$), and another matched with the peak at 13.57 min, as standard for epoxystrictosidine ($C_{27}H_{34}N_2O_{10}$) (Figure 2d). Notably, the key intermediate strictosidine was detected. The MS/MS spectrum of the ion at *m*/*z* 531 (Figure 2g) displayed the same fragmentation pattern as the molecules at: *m*/*z* 369 ($C_{21}H_{24}N_2O_8$) (*m*/*z* 531 → 369, loss of glucose with 162 Da); *m*/*z* 514 ($C_{27}H_{31}NO_9$) (*m*/*z* 531 → 514, loss of NH_3 with 17 Da); and *m*/*z* 352 ($C_{21}H_{21}NO_8$) (m/z 531 → m/z 352, loss of glucose and NH_3). This identical fragmentation pattern is consistent with the MIA strictosidine previously reported in *Strychnos peckii* (Santos et al., 2020).

Screening of *Neolamarckia* strictosidine synthase gene candidates and functional identification of NcSTR1

Previous orthogene-based analysis showed that the copy number of the STR family expanded in two Rubiaceae species (35 and 28 copies in *N. cadamba* and *Co. canephora*,

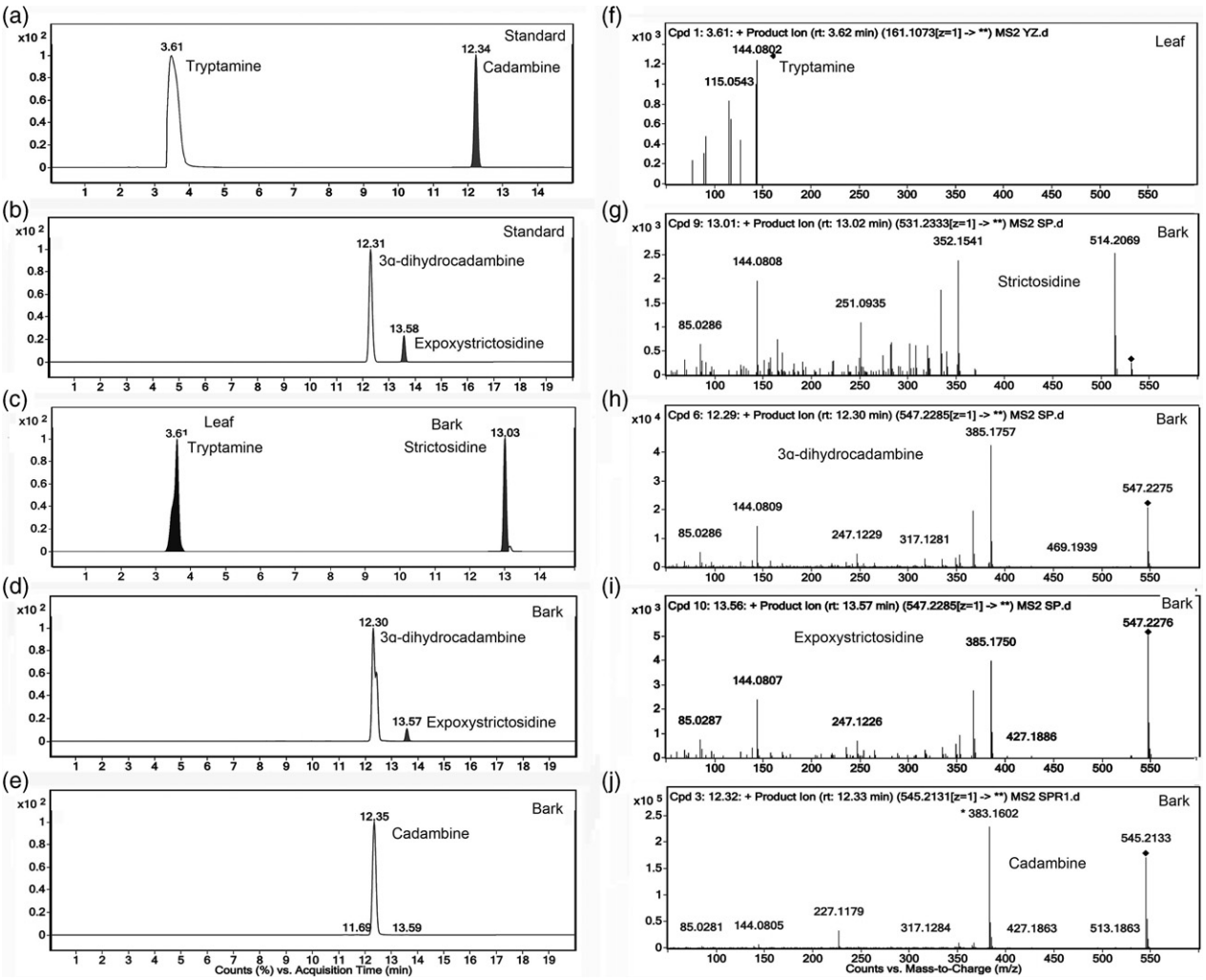

Figure 2. Components characterized by Q-TOF LC-MS/MS in *Neolamarckia cadamba*. The leaf and bark extracts of *N. cadamba* were characterized by Q-TOF LC-MS/MS (see Experimental procedures). (a, b) LC-MS chromatograms of the four authentic standards: tryptamine, cadambine, 3α-dihydrocadambine and expoxystrictosidine, respectively. (c) Detection of tryptamine from the leaves and strictosidine from the bark of *N. cadamba* by LC-MS. (d, e) Expoxystrictosidine, 3α-dihydrocadambine and cadambine were detected in the bark of *N. cadamba* by LC-MS. (f–j) MS/MS chromatograms indicated the substances represented by the peaks in (c–e) as tryptamine (f, 3.62 min), strictosidine (g, 13.02 min), 3α-dihydrocadambine (h, 12.30 min), expoxystrictosidine (i, 13.57 min) and cadambine (j, 12.33 min), respectively.

respectively), compared with *Arabidopsis thaliana*, *Catharanthus roseus*, *Theobroma cacao*. Of the 35 putative *NcSTR*s (Figure 3b), seven *NcSTR*s were grouped with a conserved clade containing the characterized strictosidine synthase gene *CRO T006099* (Kellner et al., 2015) (Figure S4), whereas only five *NcSTR*s (*NcSTR1*, *13*, *14*, *22* and *27*) share over 45% aa identities with CrSTR. Sequence alignment of these predicted NcSTRs with CrSTR1 (Pressnitz et al., 2018), RsSTR1 (Ma et al., 2006) and OpSTR1 (Eger et al., 2020) demonstrated that most previously identified active site residues, such as Cys89/Asn91/Cys101 in RsSTR1, Trp145/Tyr147 in OpSTR1 and the essential active site glutamate residue (Glu309 in RsSTR1, Glu301 in OpSTR1 and Glu315 in CrSTR1) were highly conserved in NcSTR1, NcSTR13 and NcSTR14 (Figure 3a). Moreover, *NcSTR1* (*evm.model.Contig69.90*) physically clustered with *NcTDC2* (*evm.model.Contig69.91*) in the *N. cadamba* genome and the predicted *NcTDC2* protein shares 71 and 84% sequence identities with the functional tryptophan decarboxylases of *Ca. roseus* (De Luca and Cutler, 1987) and *O. pumila* (Yamazaki et al., 2003), respectively, and thus we examined the catalytic activities of the expressed NcSTR1 enzyme first.

The coding sequence of *NcSTR1* was expressed in *Escherichia coli* and recombinant His-tagged proteins were purified by affinity chromatography (Figure S5). The purified NcSTR1 was mixed with excess tryptamine and secologanin (Figure 3c) and a reaction product of strictosidine was produced and detected by HPLC-MS/MS. Two ion pairs, 531.10 ->352.20 (blue curve) and 531.10 ->514.25

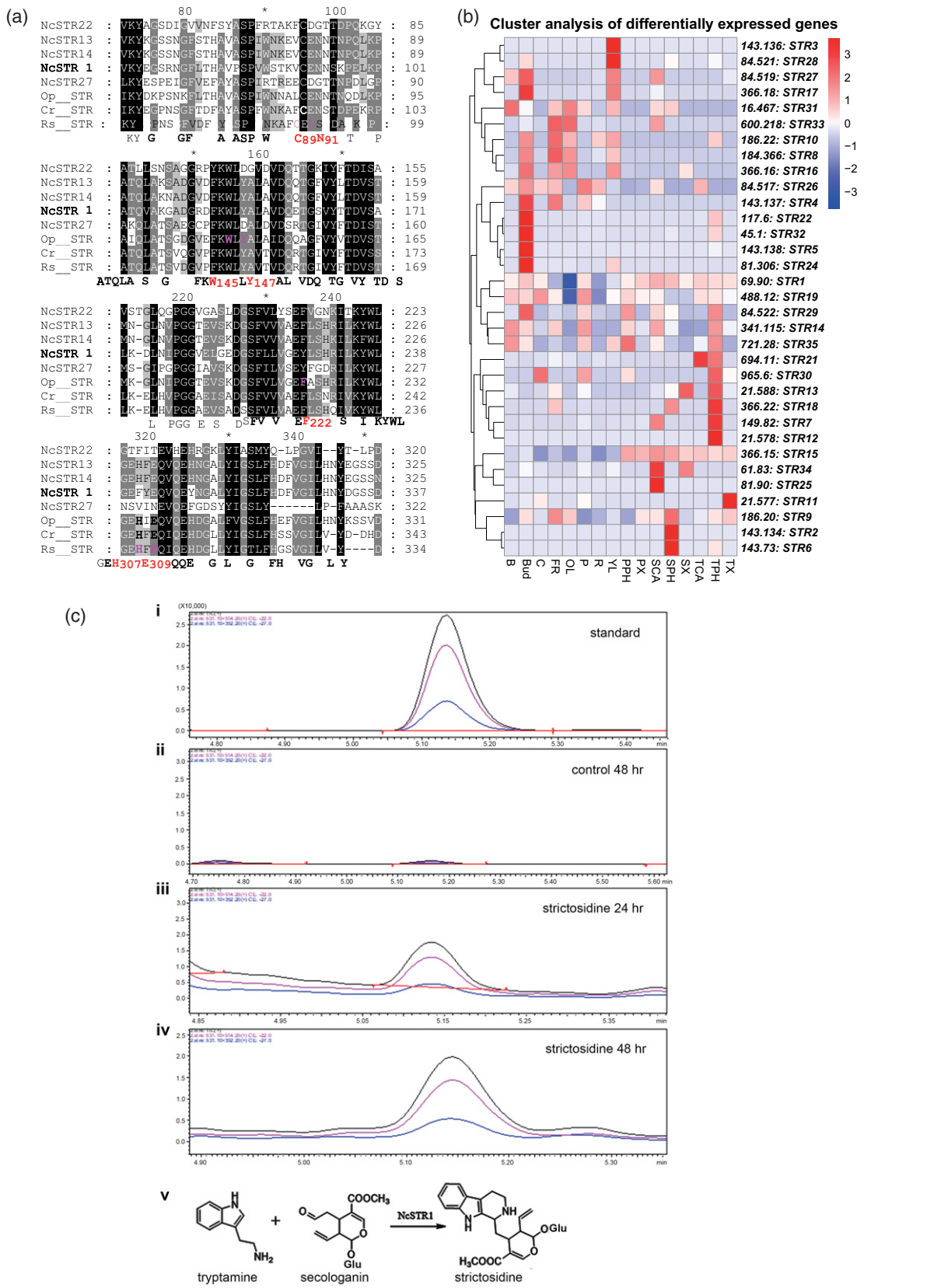

Figure 3. *Neolamarckia* strictosidine synthase gene candidate selection and functional identification of NcSTR1. (a) Alignment of the active domains in strictosidine synthases (STRs) from *Ophiorrhiza pumila* (Op), *Catharanthus roseus* (Cr), *Rauvolfia serpentine* (Rs) and the five *Neolamarckia* candidates homologous with the highest similarity to *CrSTR* at amino acid level. Previously reported active site residues of OpSTR (Q94LW9; Eger et al., 2020), CrSTR (CAA43936.1; Pressnitz et al., 2018) and RsSTR (P68175.1; Ma et al., 2006) were indicated in red letters under the aligned sequences. The essential active site glutamate residue (Glu309 in RsSTR1) was highly conserved in NcSTR1, NcSTR13 and NcSTR14. (b) The expression profile of all of the predicted 35 *Neolamarckia* STRs in 16 tissues. Bark (B), bud, cambium (C), young fruit (FR), old leaves (OL), phloem (P), root (R), young leaves (YL), xylem (primary xylem, PX; transitional xylem, TX; secondary xylem, SX), cambium (transitional cambium, TCA; secondary, SCA) and phloem (primary phloem, PPH; transitional phloem, TPH; secondary phloem, SPH) from the first, second and fourth internodes. The second internode of the 1-year-old seedling was identified as the transition from primary growth to secondary growth. (c) NcSTR1 catalyzed the Pictet–Spengler reaction of tryptamine with secologanin. In the *in vitro* catalytic reaction, tryptamine and secologanin were used as substrates to synthesize strictosidine. The curves in blue and pink indicate the 531.10 ->352.20 and 531.10 ->514.25 ion pairs, respectively; the black curve is the superimposed curve, combining the blue and pink curves. (i) The standard of strictosidine. (ii) The negative control without the enzyme NcSTR1 48 h after the reaction. (iii) The NcSTR1 catalyzed 24 h after the reaction. (iv) The NcSTR1 catalyzed 48 h after the reaction. (v) The Pictet–Spengler chemical reaction catalyzed by NcSTR1.

(pink curve), were selected as parameters for multiple reaction monitoring (MRM). With the catalyzation of NcSTR1, the peak area of the resulting strictosidine increased after 48 h of reaction, compared with 24 h (Figure 3c). The control without NcSTR1 did not have the relative peak. These results indicate that NcSTR1 could catalyze the synthesis of strictosidine in *N. cadamba*.

Integrated transcriptome and metabolome analysis for cadambine biosynthetic gene discovery

The detection of strictosidine in *Neolamarckia* bark indicated that *N. cadamba* possesses the same upstream pathway for the formation of strictosidine as that of *Ca. roseus* (Figure 4a). The gene lists involved in the mevalonate (MVA)/methylerythritol phosphate (MEP) and shikimate/indole pathways in *N. cadamba* were obtained by the preliminary screening (Table S12), and those with more than 75% aa identity to *Catharanthus* functional genes were selected as candidate biosynthetic genes (Table S13). The genes predicted to be involved in the seco-iridoid biosynthesis pathway (*GES*, *G8H*, *GOR*, *IS*, *IO*, *7-DLGT*, *7-DLH*, *LAMT* and *SLS*) in *N. cadamba* (Table S13) only contained orthologs with more than 65% aa identity with functional genes from *Ca. roseus* and *O. pumila* (Brown et al., 2015; Kai et al., 2015; Kellner et al., 2015; Salim et al., 2013). We found that most genes related to strictosidine synthesis (i.e. *GPPS*, *GES*, *DLGT*, *LAMT*, *SLS* and *TDC*) underwent the recent WGD event after diverging from *Co. canephora* (K_s < 1 in Table S14), indicating that the recent WGD event was important to the evolution of cadambine biosynthesis.

As for the downstream pathway after strictosidine synthesis in *N. cadamba*, the biogenesis of the dihydrocadambines may involve the epoxidation of strictosidine followed by the internal opening of the epoxide, closing the seven-membered ring and producing 3α-dihydrocadambine. After that, 3α-dihydrocadambine would be catalyzed by hydrolases to form cadambine (Figures 4a and S6). Consistent with this prediction, we found that the epoxystrictosidine (C27H34N2O10, *m/z* 547.2285) accumulated in *Neolamarckia* bark (Figure 2f). Therefore, the biosynthetic enzymes with monooxygenase activity, hydrolase activities, acyltransferase activity and glucosidase activities, such as cytochrome P450s (CYP), squalene epoxidases (SQE), oxidosqualene cyclases (OSCs), zeaxanthin epoxidase (ZEP), soluble epoxide hydrolase (SEH), violaxanthin de-epoxidase (VDEs), BAHD and serine-carboxypeptidase-like acyltransferases (ACT), strictosidine glucosidase (SG), etc., are proposed to serve as possible organizers in the cadambine synthetic pathway (Almeida et al., 2018; Carqueijeiro et al., 2018a, 2018b; Leonelli et al., 2017; Ma et al., 2005; Qu et al., 2018; Shang and Huang, 2020; Tatsis et al., 2017; Xia et al., 2012; Zheng et al., 2019). In order to identify more biosynthetic gene candidates, we further conducted RNA-seq analysis, GWAS, population analysis and gene cluster prediction.

RNA-seq analysis with 16 different tissues of *N. cadamba* revealed that the putative seco-iridoid biosynthetic genes had two distinctive co-expression patterns (Figure 4b). One set included *NcGES1–NcIS1-NcIO1–NcDLGT1/3*, which showed high expression in bark, bud and young leaves, and another set, *NcLAMT1/2* and *NcSLS4*, exhibited moderate expression in most of the collected tissues except for fruit, old leaves, roots and xylem. Moreover, *NcLAMT1/2* and *NcSLS4* co-expressed with *NcTDC2* (Figures 4b and S7), indicating that the synthesis of secoiridoid was coordinated with the indole precursors needed for cadambine production. In addition, *NcSTR1/14/19/29/35* and *NcSQE1/6* exhibited co-expression patterns similar to that of *NcLAMT1/2* and *NcTDC2* (Figures 3b, 4b and S8).

To assess the regulatory processes that control the accumulation of strictosidine, tryptamine and cadambine in different organs, five tissues (bark, bud, young leaves, old leaves and fruits) from the individual tree sampled for genome sequencing were used as the source of transcripts and metabolites. The transcripts were quantified by RNA-seq expression analysis (fragments per kilobase of transcript per million mapped reads, FPKM) and the metabolite contents (μg) per gram fresh weight of five tissues were determined using an LC-MS-based method. We calculated the Pearson correlations among the contents of strictosidine, tryptamine, cadambine and gene transcript levels, respectively. The correlation analysis (using a coefficient of >0.9 as the cut-off) showed that the expression level of *NcSTR1* was strongly associated with strictosidine content,

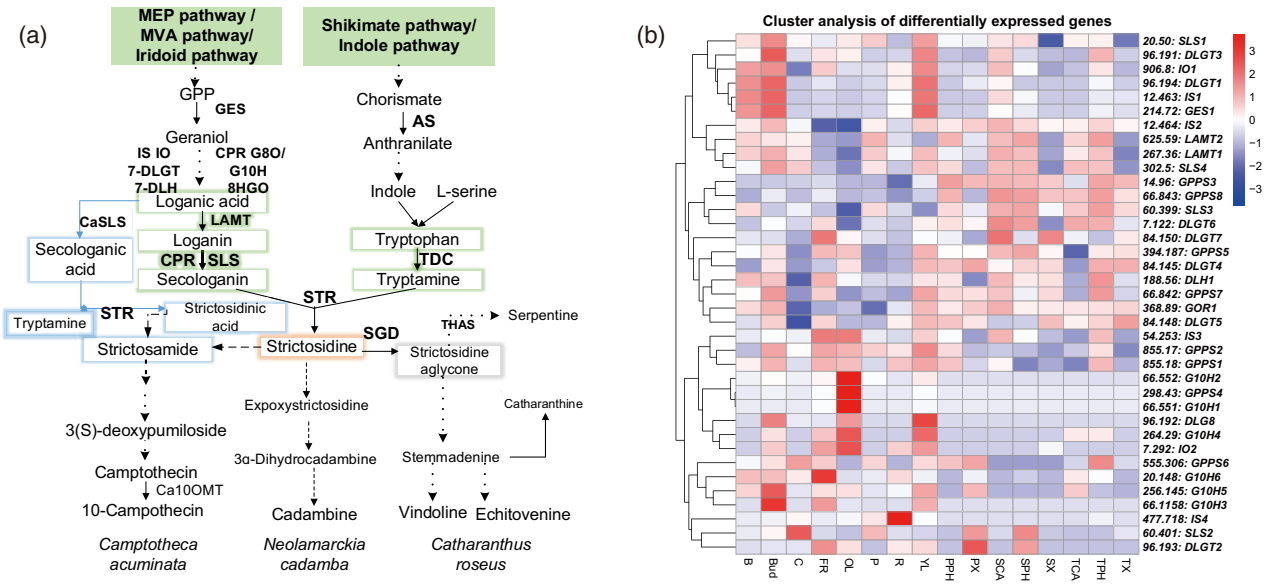

Figure 4. The predicted cadambine biosynthetic pathway and monoterpenoid indole alkaloids (MIA) biosynthetic gene candidates. (a) Proposed cadambine biosynthetic pathway in plants. Dotted lines indicate there are multiple steps between intermediates. (b) The expression profile of all *Neolamarckia* candidate biosynthetic genes for the seco-iridoid pathway in 16 tissues. Bark (B), bud, cambium (C), young fruit (FR), old leaves (OL), phloem (P), root (R), young leaves (YL), xylem (primary xylem, PX; transitional xylem, TX; secondary xylem, SX), cambium (transitional cambium, TCA; secondary, SCA) and phloem (primary phloem, PPH; transitional phloem, TPH; secondary phloem, SPH) from the first, second and fourth internodes. Enzymes in abbreviations are: 7-DLH, 7-deoxyloganic acid hydroxylase (CYP72A224); 8-HGO, 8-hydroxy-geraniol oxidoreductase; AS, anthranilic acid synthetase; Ca10OMT, *Camptotheca acuminata* 10-hydroxycamptothecin *O*-methyltransferase; CPR, NADPH-cytochrome P450 reductase; DLGT, 7-deoxyloganetic acid UDP-glucosyltransferase; G10H, geraniol-10-hydroxylase; GES, geraniol synthase; IO, iridoid oxidase (CYP76A26); IS, iridoid synthase; LAMT, loganic acid *O*-methyltransferase; SGD, strictosidine β-glucosidase; SLS, secologanin synthetase; STR, strictosidine synthase; TDC, tryptophan decarboxylase; THAS, tetrahydroalstonine synthase (Kai et al., 2015; Wu et al., 2018; Rai et al., 2021; Qu et al., 2018; Yang et al., 2019).

and that *NcSTR14* (*evm.model.Contig341.115*) and *NcSLS3* (*evm.model.Contig60.399*) were strongly associated with cadambine content (Table S15). Two *NcCYP*s (*evm.model.Contig208.74* and *evm.model.Contig708.57*) and *NcOSC1* (*evm.model.Contig207.232*) were also identified to be strongly associated with cadambine content in the network. Both NcCYPs were assigned into the CYP72A subfamily, which contained numerous essential components in MIA-producing plants (Figure S9) (Urlacher and Girhard, 2019; Zheng et al., 2019). In addition, the expression level of 12 transcription factors (TFs) were highly correlated with the cadambine accumulation pattern. Of these, *evm.model.Contig28.407* putatively encoded a basic helix-loop-helix (bHLH) TF with 50 and 45% aa identity with *CrBIS1* and *CrBIS2*, respectively (Van Moerkercke et al., 2015, 2016).

GWAS and population identification of genes potentially related to cadambine biosynthesis

To gain further insights into the cadambine synthetic pathway, we conducted a GWAS using 112 individuals collected from 27 populations distributed widely in Southeast Asia (Table S16). We identified 5 786 667 high-quality single-nucleotide polymorphisms (SNPs) based on the reference genome. To establish the SNP-based identification of genotype–phenotype associations, we examined the contents of strictosidine, tryptamine and cadambine in the extracts from the bark and young leaves of 112 *Neolamarckia* accessions by high-resolution mass spectrometry. The association between each SNP and the phenotype was conducted in a mixed linear model using GEMMA (Zhou and Stephens, 2014). GWAS revealed that the candidate genomic loci responsible for *Neolamarckia* strictosidine and cadambine accumulation were associated with *NcSTR3/4/13*, *NcTDC2* and most of the putative genes in the seco-iridoid biosynthesis pathway (Table S17). GWAS also revealed new biosynthetic gene candidates, including *NcSQE7*, *NcSEH7*, *NcCPR1* and six *NcCYP*s, putatively involved in the epoxidation and hydrolysis of epoxide (Meijer et al., 1993; Zheng et al., 2019), that were associated with the significant loci responsible for cadambine accumulation (Table S18).

In this study, we also conducted population-level analysis of genetic variation of *N. cadamba* using 31 accessions with high, medium or low cadambine levels. To observe the divergence among the three groups at the genomic level, we constructed a neighbor-joining phylogenetic tree and performed principal component analysis (PCA) of members of clade 3 (Figure 5a,b). We observed similar results in phylogenetic tree and PCA analyses.

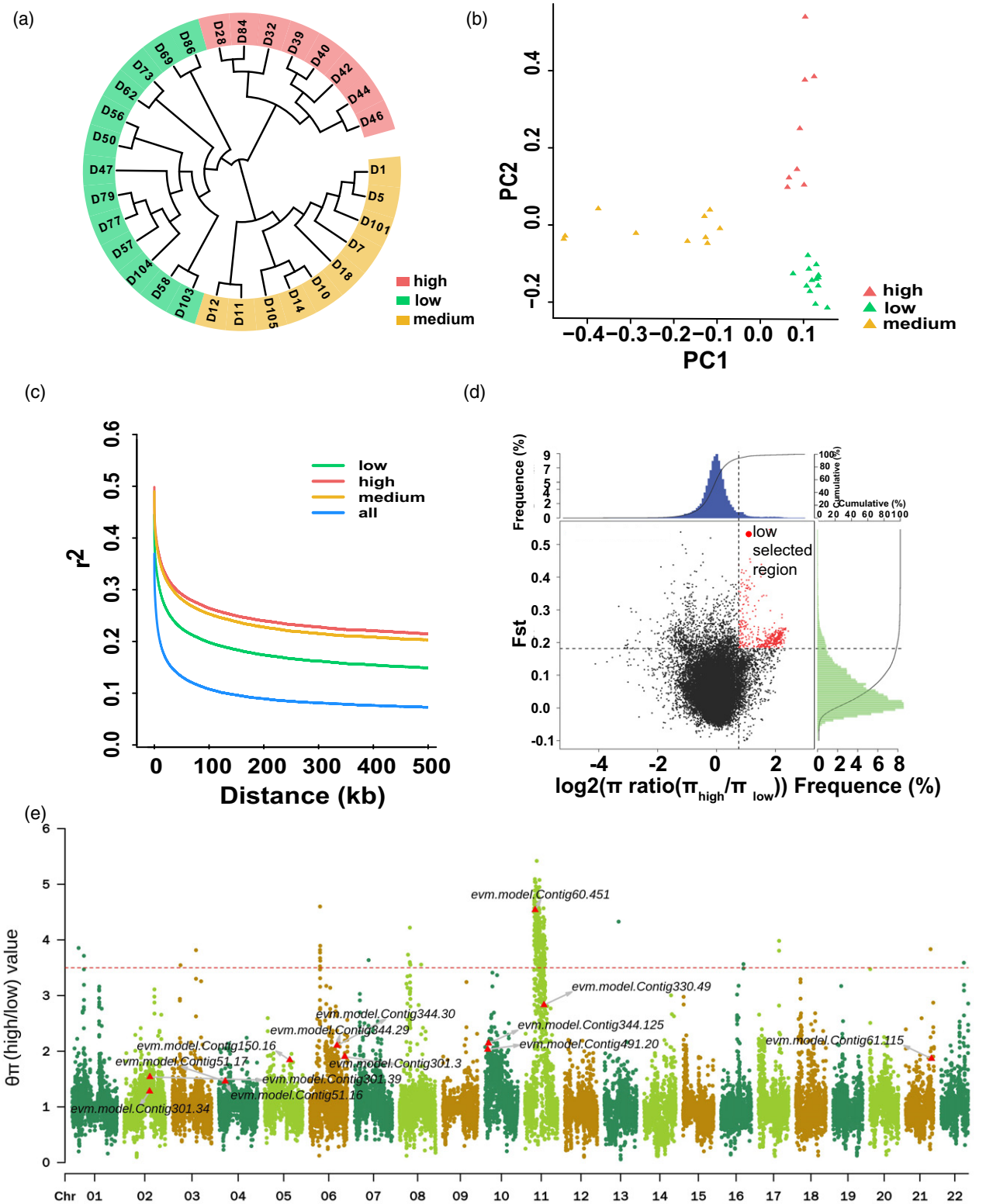

Figure 5. Population analysis of *Neolamarckia cadamba* accessions based on cadambine content. (a) A neighbor-joining phylogenetic tree of the 31 accessions based on their cadambine content. (b) Principal component analysis of the selected 31 accessions. (c) Decay of linkage disequilibrium (LD), measured by r^2, in the four groups. (d) Selective signals in the whole genome between different ecotypes. (e) Manhattan plot of θπ-based detection of selective sweeps identified by comparison between two groups with high and low cadambine content. KEGG enrichment candidate genes associated with cadambine biosynthesis are highlighted with red arrows.

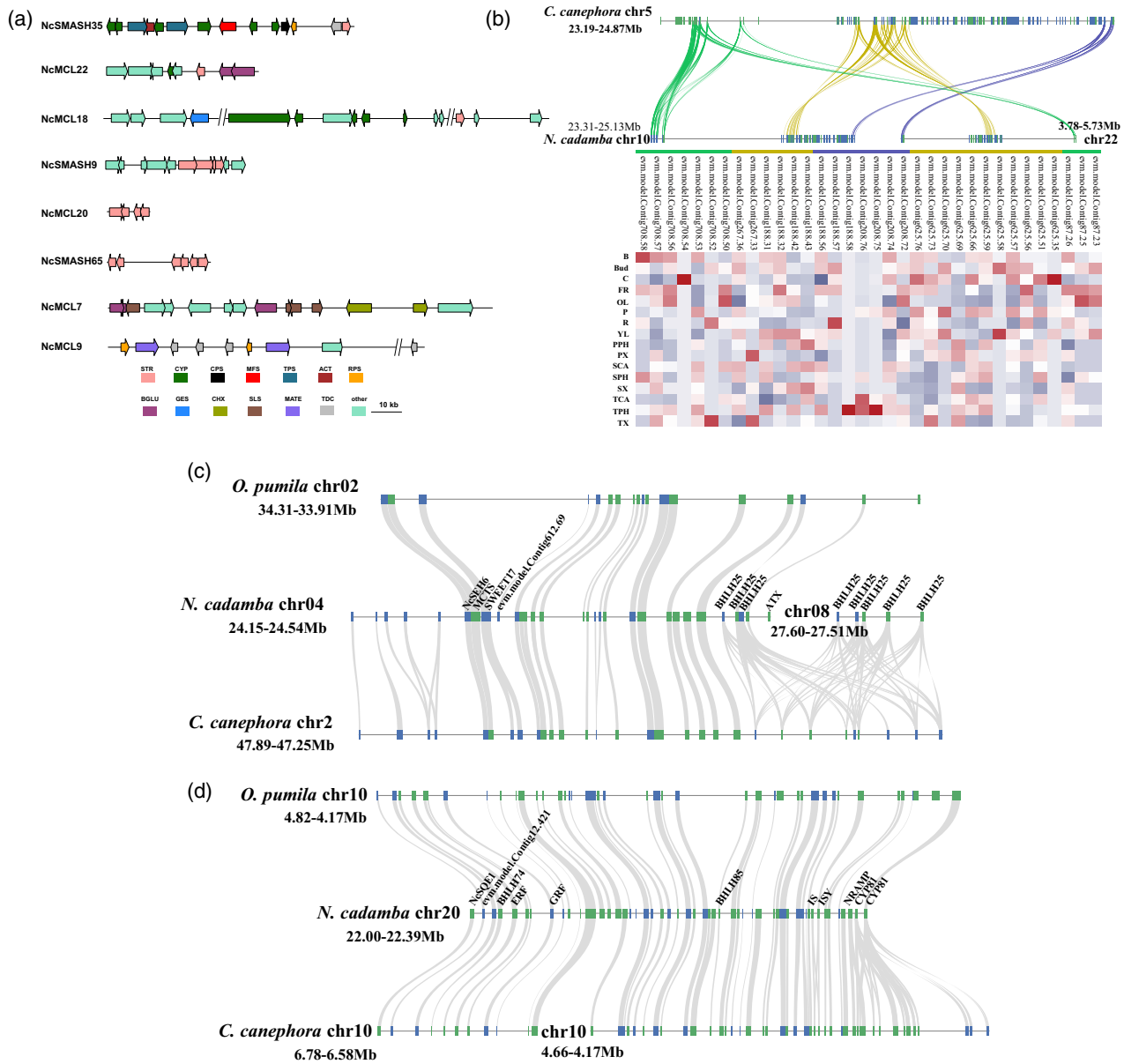

Figure 6. The structure and evolution of the predicted cadambine biosynthetic gene clusters and tandem duplications. (a) Schematic diagram of the predicted gene clusters and tandem duplicates in *Neolamarckia cadamba*: NcSMASH35, TDC/STR1/CYPs; NcMCL22, CYP716/NcSTR14/BGLUs; NcMCL18, GES/NcSTR13/CYPs; NcSMASH9, NcSMASH65 and NcMCL20, tandem duplicated NcSTRs; NcMCL7, tandem duplicated NcSLSs; NcMCL9, tandem duplicated NcTDCs. (b) Evolution of two genomic regions encompassing gene clusters NcMCL10, NCMCL19, NcMCL23 and NcSMASH67 and the expression profile of the putative biosynthetic genes in these clusters: NcMCL10, NcLAMT1/NcDL7H/CYPs; NcMCL19, NcLAMT2/Cytb5/CYP; NcMCL23 and NcSMASH67, tandem duplicated NcCYP72s. (c) Evolution of the predicted gene cluster with NcSQE1/IS/ISY/CYP81. (d) Evolution of tandem duplicated basic helix-loop-helix (bHLH) transcription factors (TFs) in NcMCL25 and NcMCL24. Enzyme abbreviations: ACT, acyltransferase; ATX, copper transport protein ATX family; BGLU, β-glucosidase; BHLH25, transcription factor bHLH25; CHX, cation/H(+) antiporter; CPS, terpenoid cyclases; CYP, cytochrome P450s; ERF, ethylene-responsive transcription factor; GES, geraniol synthase; GRF, growth-regulating factor; IS, iridoid synthase; ISY, iridoid synthase paralog; MATE, multidrug and toxin extrusion protein; MCTS, malignant T-cell-amplified sequence; MFS, major facilitator superfamily protein; NRAMP, metal transporter Nramp family; PRS, disease resistance protein; SEH, soluble epoxide hydrolase; SLS, secologanin synthetase; SQE, squalene epoxidase; SWEET, bidirectional sugar transporter; TDC, tryptophan decarboxylase; TPS, terpenoid synthase.

We also analyzed the linkage disequilibrium (LD) throughout the whole genome. LD (indicated by r^2) decreased with physical distance between SNPs in all groups. The average distance of LD for each group was measured as the chromosomal distance when LD decreased to half of its maximum value. The three groups showed different extents of genome-wide LD decay, with LD decaying fastest in populations with low cadambine levels and slowest in populations with high cadambine levels (Figure 5c). Then, we explored the genomic regions

with high divergence to try to identify genes possibly involved in cadambine metabolism and/or accumulation. According to an empirical procedure described in a previous study (Li et al., 2013), the intersection regions with the top low or high π ratios and the top high fixation index (F_{ST}) values between ecotype groups were identified as selective sweeps.

We calculated the F_{ST} values between subgroups, and the genomic regions with F_{ST} values in the top 5% were considered highly differentiated (Weir and Cockerham, 1984). We estimated the population diversification parameters, π and θ_W, and found that the overall nucleotide diversity in species with high cadambine level was higher than that in species with low or medium cadambine levels (high-cadambine level species, $\pi = 0.0023440$ and $\theta_W = 0.0019451$; low-cadambine level species, $\pi = 0.0023134$ and $\theta_W = 0.0018883$) (Figure 5d). KEGG analysis (FDR < 0.05) indicated the genes harbored in these selective sweeps were involved in cadambine biosynthesis (e.g. sesquiterpenoid and triterpenoid biosynthesis, terpenoid backbone biosynthesis and indole alkaloid biosynthesis; Figure 5e; Table S19).

Prediction of biosynthetic gene clustering and tandem duplication in *N. cadamba*

Biosynthetic gene clusters (BGCs) for a wide variety of natural products have now been reported from diverse plant species (Li et al., 2021). The definition of a metabolic gene cluster requires that it should contain genes encoding at least three different types of tailoring enzymes (Jacobowitz and Weng, 2020; Nützmann et al., 2016; Schläpfer et al., 2017). We systematically mined the *N. cadamba* genome sequence to identify all the BGCs with the plantiSMASH algorithm (Kautsar et al., 2017). This identified 67 possible gene clusters across the 22 pseudochromosomes (Table S20). Moreover, we identified additional BGCs that included at least one gene involved in the seco-iridoid pathway, or *NcTDC/NcSTR*, with a physical size of <600 kb (Table S21).

A notable 12-gene cluster (*NcSMASH35*) within a 340-kb region in chromosome 11 was identified, which included functionally characterized *NcSTR1*, *NcTDC2* and five tandem duplicated *CYP71D* subfamily members, a *CYP76* member and four *TPS*s (Figure 6a). A predicted transporter gene *evm.model.Contig69.96* putatively encoding a major facilitator superfamily protein with moderate expression in all the tissues was present in this *NcTDC2*/*NcSTR1* gene cluster, similar to the previous reports in *Rhazya stricta* (Sabir et al., 2016), *Gelsemium sempervirens* and *O. pumila* (Rai et al., 2021). All the six *NcCYP*s in this cluster were derived from the recent WGD event ($K_s < 1$), and the closest match of the five *NcCYP71D* paralogs is the *Ca. roseus* tabersonine 3-oxygenase gene (*CrT3O*, AEX07771) with 51–58% identity at the amino acid level (Figure S9).

In *N. cadamba*, we identified two genes with 88% identity at the nucleotide level (*evm.model.Contig267.36* and *evm.model.Contig625.59*) putatively encoding loganic acid *O*-methyltransferase (LAMT). The sequence alignment of the predicted NcLAMT1/2 with CrLAMT (CRO_T028497) revealed that all the identified active site residues of CrLAMT (Y159, H162, W163, P227, A241, H245, Q273, H275, P302, Q316, I320 and D359) were conserved in NcLAMT1/2 (Figure S10). Two genomic regions with NcLAMT1/2 located on pseudochromosome 10 (25.31–25.13 Mb) and pseudochromosome 22 (3.78–5.73 Mb) were identified, and included the gene clusters *NcMCL10*, *NCMCL19*, *NcMCL23* and *NcSMASH67* (Figure 6b). In NcMCL10, the *Neolamarckia*-specific 7-DLH homolog *evm.model.Contig188.56* (86% aa identity with *CrDL7H*) is located approximately 600 kb from *NcLAMT1*, and two other *NcCYP72A*s were closely adjacent to the predicted *NcDL7H*. Interestingly, the two cadambine biosynthetic candidates (*evm.model.Contig208.74* and *evm.model.Contig708.57*) identified in the gene–metabolite network were also distributed in *NcSMASH67* and *NcMCL23*, respectively, implying a selection pressure favoring the clustering of genes associated with MIA production in *N. cadamba*.

The gene cluster *NcMCL15* located on pseudochromosome 20 consists of *NcIS1/2*, *NcSQE1* and two *NcCYP81*s (Figure 6c). A syntenic region with high sequence similarity and gene localization on pseudochromosome 16 was identified in *N. cadamba*. Synteny analysis among *O. pumila*, *N. cadamba* and *Co. canephora* further revealed that *N. cadamba* had undergone a WGD. This is further supported by the low median K_s value (approx. 0.41) of gene paralogs for most genes in *NcMCL15* after excluding the two key genes *NcIS1* and *NcIS2*. *NcSQE1* is located approximately 300 kb from *NcIS1* in cluster *NcMCL15*, and the co-expression profiles of *NcSQE1/6*, *NcLAMT1/2*, *NcTDC2* and *NcSTR1/14* suggest that *NcSQE1/6* might be involved in the epoxidation of strictosidine.

Another predicted cluster (*NcSMASH17*) in *N. cadamba* is located on pseudochromosome 4 and consists of *NcSQE4*, an MYB TF gene with high similarity to *OpMYB1* (BAU61355.1, 83% aa identity) and a subtilisin-like protease (*SBT*) gene associated with cadambine content in the gene–metabolite network (Table S15). *OpMYB1* is an R2R3-MYB repressor that acts as a negative regulator of MIA production, and its overexpression in hairy roots of *O. pumila* resulted in the reduced production of camptothecin and the reduced expression of OpTDC (Ma and Constabel, 2019; Rosseleena et al., 2016).

Previous studies demonstrated that specialized metabolic genes were more significantly enriched in local (tandem) duplication events as compared with WGD events (Chae et al., 2014). We noted that in the *N. cadamba* genome the presence of multiple paralogs of several

predicted MIA biosynthetic genes putatively encoding 7-DLGT, IS, ISY, TDC, SLS, STR and CYPs were also significantly enriched in local (tandem) duplication events (Figures 6a,b and S11).

In addition, tandem duplication of TFs involved in specialized metabolite biosynthesis were also detected in a wide range of eudicots (Colinas and Goossens, 2018). In *Ca. roseus*, three bHLH iridoid synthesis gene (*BIS*) tandem duplicates (*CrBIS1/2/3*) exclusively transactivated the expression of the biosynthetic genes in the seco-iridoid pathway (Van Moerkercke et al., 2015, 2016; Singh et al., 2021). In the gene–metabolite network, we characterized one homolog as *CrBIS1* (*evm.model.Contig28.407*). Analysis of the genome sequence of *N. cadamba* showed that four additional bHLH TF genes clustered with *evm.model.Contig28.407* in a tandem duplication within a 130-kb region on pseudochromosome 4 (*NcMCL24*, Figure 6d). Synteny analysis further detected another bHLH TF tandem duplicate (*NcMCL25*) with close homology and gene localization with *NcMCL24* on pseudochromosome 8, which includes significant loci responsible for strictosidine accumulation in GWAS (24.25–24.28 Mb). The co-expression profile of some bHLH TFs in *NcMCL24* and *NcMCL25* suggest co-regulation mechanisms in specialized metabolite biosynthesis (Figure S12). In *NcMCL26* (Figure S13), four NcMYC TFs were clustered in tandem order within a 24-kb region, and these loci were found to be responsible for strictosidine accumulation in GWAS. Of the four MYC TFs, *evm.TU.Contig184.722* showed the highest similarity (77% for the deduced amino acid sequence) with *CrMYC2* (AF283507), the major activator of MeJA-responsive *ORCA2/3*, which in turn regulated a subset of alkaloid biosynthesis genes in *Ca. roseus* (Zhang et al., 2011).

DISCUSSION

Historically (Mehra and Bawa, 1969), and in the Chromosome Counts Database (http://ccdb.tau.ac.il/home), the *Anthocephalus cadamba* (Roxb.) Miq (now accepted as *N. cadamba* (Roxb.) Bosser) gametophytic chromosome number *n* was 22. *Neolamarckia cadamba* was characterized as a tetraploid species in those relatively old references (Bedi et al., 1981; Mehra and Bawa, 1969). As there were confusing names for *N. cadamba* (Pandey and Negi, 2016), first we identified that the sequenced *N. cadamba* is diploid, with $2n = 2x = 44$ chromosomes, with chromosome measurements and a pair of *45S* rDNA probe signals with fluorescent *in situ* hybridization (FISH) (Figure S1). This provided a solid base for establishing high-quality *N. cadamba* Hi-C libraries.

To date, three species in the genus *Coffea* (*Coffea arabica*, *Coffea canephora* and *Coffea eugenioides*), *O. pumila* and *Gardenia jasminoides* in the family Rubiaceae have been sequenced. *Neolamarckia cadamba* is the first sequenced tree species in this family. The genus *Neolamarckia* is a ditypic genus that included only two tree species: *N. cadamba* and *Neolamarckia macrophyllus* (endemic to Sulawesi, Indonesia; Li et al., 2018). This genus features densely globe-shaped flower clusters (from which the Chinese name 'TuanHua' originates) with an orange scent (Kareti and Subash, 2020). Another representative characteristic of this genus is its rapid growth and remarkable canopy, with a height of 45 m and a stem diameter of 100–160 cm. Notably, *N. cadamba* is the only species reported so far that accumulates a high content (approx. 0.1%) of bioactive 3-dihydrocadambine and cadambine in the bark and leaves. Given that we have established a highly efficient regeneration system of *N. cadamba* (Li et al., 2019), the genomic information presented here will greatly facilitate the elucidation of the cadambine synthetic pathway and the development of tools for enhancing bioactive productivity by metabolic engineering in microbes or by molecular breeding in plants.

Furthermore, comparative genomic analysis among *Co. canephora*, *N. cadamba* and *O. pumila* will shed light on the evolutionary history of specific metabolic pathways, as they produce quite different major specialized metabolites, including cadambine, caffeine (a purine alkaloid) and camptothecin (anti-cancer MIAs). The emergence of STR for strictosidine synthesis was generally considered an important innovative step for the strictosidine-derived MIA-producing plants (Rai et al., 2021). However, no candidates homologous (aa identity cut-off value of >55%) with the late seco-iridoid pathway genes *NcLAMT1*/*NcLAMT2* and *NcSLS1-4* were identified in the *Co. canephora* genome. In contrast to *Coffea*, the *Ophiorrhiza* homologous candidates *OpLAMT* (Opuchr05_g0056110) and *OpSLS*s (Opuchr02_g0012990, Opuchr02_g0013060, Opuchr02_g0013090 and Opuchr02_g0017930) share 84% aa identity with *NcLAMT1/NcLAMT2* and 67–85% aa identity with *NcSLS1-4*, respectively. *Neolamarckia cadamba* diverged approximately 41.7 mya from *O. pumila* and approximately 31.0 mya from *Co. canephora* (Figure 1a), we therefore propose that the gene evolution of *LAMT* involved a deletion following a duplication process, and that the duplication of *NcLAMT* derived from a recent WGD event was key for the evolution of cadambine biosynthesies in *N. cadamba*, as there is only one copy of the *LAMT* gene in the well-known MIA-producing species *Ca. roseus* and *O. pumila* (Figure S10). Moreover, although camptothecin is found in *O. pumila*, OpLAMT, OpSLS and OpSTR have the same functions as those of *Ca. roseus* and similarly produce loganin, secologanin and strictosidine. Rai et al. (2021) also observed that the three key *N*-methyltransferases (NMTs), xanthosine methyltransferase (CcXMT), theobromine synthase (7-methylxanthine methyltransferase, CcMXMT) and caffeine synthase (3,7-dimethylxanthine methyltransferase, CcDXMT), essential for caffeine biosynthesis are missing in the *N. cadamba*

genome, as the key residues Gln161, Ile266, Ser316 and Tyr356 for CcXMT1 (Mccarthy and Mccarthy, 2007) were no longer conserved in the deduced amino acid sequences of the *Neolamarckia* homologs, providing a reasonable explanation for why *N. cadamba* lacks the first methylation steps necessary to produce caffeine from xanthosine (Figure S10).

Two *O*-β-D-glucosidases, strictosidine *O*-β-D-glucosidase (SG, EC3.2.1.105) and raucaffricine *O*-β-D-glucosidase (RG, EC3.2.1.125), act as major components in the MIA biosynthesis pathway. In *Ca. roseus* and *Rauvolfia*, SG follows strictosidine synthase in the production of the reactive intermediate required for the diverse MIAs, whereas *Rauvolfia* RG hydrolyzes the glucoalkaloid raucaffricine, forming the aglycone vomilenine, an intermediate that appears in the middle of the ajmaline pathway (Barleben et al., 2007; Xia et al., 2012). From a biosynthetic point of view, it is extremely rare for a glucoside such as the glucoalkaloid strictosidine to act as a precursor at the beginning of the biosynthetic pathway and for it to become activated by deglucosylation (Barleben et al., 2007). We therefore explored whether there are similar or identical enzymes to SGs or RsRG in *N. cadamba*. Homology searching and protein sequence alignment showed that all *Neolamarckia* and *Ophiorrhiza* homologs of CrSG (Rai et al., 2021) were highly conserved in the essential active site residues (Glu207, Glu416, His161 and Trp388) of RsSG (Barleben et al., 2007), whereas they were not conserved at four critical active site residues (Thr189, His193, Tyr200 and Ser390) of RsRG (Figure S14), suggesting that these *Neolamarckia* genes are more likely assigned to plant SGs. This analysis suggested that some members of strictosidine *O*-β-D-glucosidases probably play a role in the production of the reactive intermediate of the diverse MIAs in *N. cadamba*.

The knowledge about selective sweeps provides insights and targets for the use of germplasm. Moreover, new genetic variation is needed to increase the cadambine content. In our study, we reported the genome variation mapping of 112 accessions and detected numerous diverse selective sweeps among ecotype groups in association with environmental adaptability and cadambine-related traits by analyzing genome structure diversifications between the three ecotype groups. The information from selective sweeps associated with agriculturally important markers will be helpful for molecular breeding in *N. cadamba*.

EXPERIMENTAL PROCEDURES

Chromosome preparation and FISH analysis

For the improved characterization of the chromosomes, *45S* rDNA was labeled with Chroma Tide Alexa Fluor 488–5-dUTP (Invitrogen, now ThermoFisher Scientific, https://www.thermofisher.com) for FISH and the mitotic metaphase spreads were prepared from meristem root tip cells of *N. cadamba* following the procedures described by Deng et al. (2012), with minor modifications.

Plant materials, DNA extraction and genome sequencing

Young tender leaves of *N. cadamba* collected from a 7-year-old individual plant in South China Agricultural University (Guangzhou, China) were used for DNA extraction. A paired-end (PE) library (insert size 350 bp) was constructed and sequenced on the Illumina Xten platform (https://www.illumina.com).

SMRTBell libraries with an insert size of 20 kb were also constructed and sequenced on the PacBio Sequel platform (https://www.pacb.com). For 10X Genomics sequencing, DNA sample preparation, indexing and barcoding were performed using the GemCode Instrument (10X Genomics, https://www.10xgenomics.com).

Genome assembly and assessment

De novo assembly of the PacBio reads was performed using FALCON (https://github.com/PacificBiosciences/FALCON) and FALCON-UNZIP. With the phasing information from the raw reads, this generates a subsequent set of primary contigs and these contig sequences were polished using QUIVER (http://pbsmrtpipe.readthedocs.io/en/master/getting_started.html). The 184.45-Gb 10X Genomics data were aligned with the initial assembly using BWA and the scaffolding approach was performed with FRAGSCAFF. After that, we used PBJELLY to fill gaps with PacBio data, with the parameters: -minMatch 8 -sdpTupleSize 8 -minPctIdentity 75 -bestn 1 -nCandidates 10 -maxScore -500 -nproc 13 -noSplitSubreads. Finally, the anchorage of the genome assembly onto chromosomes was performed with the LACHESIS pipeline.

To evaluate the completeness and quality of the assembly, high-quality reads from short insert sizes were mapped to the assembly using BWA (Li and Durbin, 2009). Additionally, CEGMA (CORE EUKARYOTIC GENES MAPPING APPROACH) defined a set of conserved protein families that occur in a wide range of eukaryotes and identified their exon–intron structures in genomic sequences. We also used BUSCO (BENCHMARKING UNIVERSAL SINGLE-COPY ORTHOLOGS) to assess the completeness of the genome assembly.

Repetitive elements and genes annotation

Homology was used to search the *de novo* assembly for transposable elements (TEs) in the *N. cadamba* genome. For this approach, REPEATMODELER (http://www.repeatmasker.org/RepeatModeler/), REPEATSCOUT, LTR-FINDER and TANDEM REPEATS FINDER (TRF) were used. For homology prediction, REPEATMASKER 3.3.0 and REPEATPROTEINMASK were used against the Repbase TE library (http://www.girinst.org/repbase) and the TE protein database, respectively.

Three independent approaches, including homology alignment, *de novo* search and transcriptome prediction was applied to predict protein coding genes in the *N. cadamba* genome.

Homology-based gene prediction: homolog protein sequences of *Arabidopsis thaliana*, *Coffea canephora*, *Populus trichocarpa* and *Solanum tuberosum* were downloaded from Ensemble (http://plants.ensembl.org/index.html) and NCBI (https://www.ncbi.nlm.nih.gov) and then aligned with the *N. cadamba* genome assembly using tblastn (E-value $1e^{-5}$).

Ab initio gene prediction: AUGUSTUS 2.5.5, GENSCAN 1.0, GLIMMERHMM 3.0.1, GENEID and SNAP were used to predict coding regions in the repeat-masked genome.

Transcriptome-assisted gene prediction: TOPHAT 2.0.8 (Kim et al., 2013; Trapnell et al., 2009) was used to map clean RNA-seq reads to the *N. cadamba* genome and CUFFLINKS 2.1.1 (http://cufflinks.

cbcb.umd.edu; Trapnell et al., 2012) was then used to assemble the transcripts into gene models (Cufflinks-set).

Gene model evidence from homology, *ab initio* and CUFFLINKS prediction were combined by EVIDENCEMODELER (EVM, http://evidencemodeler.github.io) into a non-redundant set of gene structures.

Functional annotation protein coding genes

The predicted protein sequences were searched against six protein/function databases: InterPro, Pfam, SwissProt, NR, GO and KEGG. Searches of the InterPro and Pfam databases were performed using INTERPROSCAN 4.8 and HMMER 3.1, respectively. For the other databases, BLAST searches were performed with an E-value cut-off of $1e^{-5}$.

Species phylogenetic analysis

ORTHOMCL (http://orthomcl.org/orthomcl) was used to cluster paralogs and orthologs among 14 species (*Actinidia chinensis*, *A. thaliana*, *Co. canephora*, *Elaeis guineensis*, *Eucalyptus grandis*, *Glycyrrhiza uralensis*, *Malus domestica*, *N. cadamba*, *O. pumila*, *Oryza sativa*, *P. trichocarpa*, *S. tuberosum*, *Theobroma cacao* and *Vitis vinifera*). We obtained the similarity relationships between protein sequences for all species through all-vs-all blastp with an E-value of $1e^{-5}$. MUSCLE (http://www.drive5.com/muscle/; Edgar, 2004) was used to align all 402 single-copy gene protein sequences and make a super alignment matrix. Then, RAXML (http://sco.h-its.org/exelixis/web/software/raxml/index.html; Stamatakis, 2014) was used to construct the 14-species phylogenetic tree using the maximum-likelihood method. Finally, MCMCTREE (http://abacus.gene.ucl.ac.uk/software/paml.html; Puttick, 2019) was applied to infer the divergence time based on the phylogenetic tree constructed. CAFÉ 2.2 (De Bie et al., 2006) was used to determine the expansion and contractions of orthologous gene families.

Whole-genome duplication (WGD)

To identity syntenic blocks, protein sequences from *Co. canephora*, *N. cadamba* and *S. tuberosum* were searched against themselves using blastp (E-value, $<1e^{-5}$) (Scott and Madden, 2004). Syntenic blocks were determined by MCSCANX (Wang et al., 2012) with the parameter of at least five genes per block. We then calculated the 4DTv (fourfold degenerate synonymous sites of the third codons) distances for syntenic segments from the concatenated alignments, which were used to construct fourfold degenerate sites of all gene pairs found in each segment and the distribution of the 4DTv values was plotted. The synonymous substitution rate (K_s) was also calculated using the MYN algorithm (Wang et al., 2009), based on the Tamura–Nei model to further confirm the most recent WGD.

NMR and LC-MS/MS analyses

Fresh leaves and bark collected from the same tree for genome sequencing were ground and extracted with 70% aqueous ethanol (v/v) for 20 min. The extract was prepared for chromatograph analysis. The NMR spectra of cadambine from Song's Laboratory (Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences) was characterized on a Bruker AV400 spectrometer (https://www.bruker.com) using tetramethylsilane (TMS, $\delta = 0$) as the internal reference. The spectra were consistent with the data for the reference (Figure S3; Yuan et al., 2020).

All the intermediates were detected by UPLC-Q-TOF (UPLC1290-6540B Q-TOF; Agilent, https://www.agilent.com). The mobile phase comprised acetonitrile (A) and 0.2% formic acid + 10 mM ammonium formate (B). The synthesized strictosidine *in vitro* catalyzed by NcSTR1 was analyzed by HPLC-MS/MS (LCMS-8050; Shimadzu, https://www.shimadzu.com). A binary gradient elution with a flow rate of 0.3 ml min^{-1} was performed as follows: 6 min, 25% methanol; 5 min, 90% methanol; 5.1 min, 10% methanol; and 7 min, 10% methanol. The temperature of the column oven is 40°C and the temperature of the sample tray is 4°C. The standard substances included secologanin (19351-63-4; Absin, https://www.absin-bio.com), tryptamine (61-54-1; Sigma-Aldrich, https://www.sigmaaldrich.com), and 3α-dihydrocadambine and epoxystrictosidine (54483-84-0; Nature Standard, http://www.naturestandard.com).

The *in vitro* catalytic assay of the recombinant strictosidine synthase NcSTR1

The coding sequence of *evm.model.contig69.90* without the start codon was amplified with a primer pair (forward primer, 5′-CTCTTGGAAATCATGCCTCACA-3′; reverse primer, 5′-TCAGA CAGAAGAAACCACTCCATTC-3′) and cloned into the pEASY®-Blunt E1 Expression Vector (TransGen Biotech, https://www.transgenbiotech.com), and further transformed in *E. coli* DH5α. The recombinant plasmids were then introduced into *E. coli* BL21 Rosetta DE3 strains for NcSTR1 protein expression. The recombinant NcSTR1 production and purification was obtained as the procedure described in the reference.

The procedure of *in vitro* catalytic assay was modified from Pressnitz et al. (2018). Generally, the catalytic reaction solution consisted of purified enzyme preparation (10 μg) dissolved in PIPES buffer (50 mM, pH 6.1), 5 mM secologanin and 5 mM tryptamine, with a total volume of 1000 μl. The mixtures were incubated on a shaker at 35°C for a given time (0, 24, 48 h). The reaction was terminated by the addition of 10 M NaOH (100 μl). Then, ethyl acetate (500 μl) was added to precipitate the PIPES. The collected supernatant was dried by nitrogen gas and then dissolved in 100 μl 50% methanol and subjected to HPLC-MS/MS analysis.

RNA-seq analysis

To capture diverse gene expression, we extracted RNA from 16 tissues of *N. cadamba*, namely bark (B), bud, cambium (C), young fruit (FR), old leaves (OL), phloem (P), root (R), young leaves (YL), xylem (primary xylem, PX; transitional xylem, TX; secondary xylem, SX), cambium (transitional cambium, TCA; secondary, SCA) and phloem (primary phloem, PPH; transitional phloem, TPH; secondary phloem, SPH), from the first, second and fourth internodes following the RNeasy Plant Mini Kit protocol (Qiagen, https://www.qiagen.com). Clean RNA-seq reads were mapped to the *N. cadamba* reference genome using HISAT2, and then the expression level for *N. cadamba* genes (FPKM, and expression count data) was obtained using HTSEQ (Kim et al., 2015b, 2019).

Identifying genomic regions responsible for cadambine accumulation by GWAS

The strictosidine and cadambine contents of bark and leaves from 112 individuals were determined by UPLC-Q-TOF (UPLC1290-6540B Q-TOF; Agilent), as described above (LC-MS/MS). GWAS analyses were performed in the mixed linear model using GEMMA 0.94.1 (Zhou and Stephens, 2012). The Bonferroni correction was used to reduce the probability of false positives, and $P = 10^{-5}$ was used as the genome-wide threshold to screen for significantly associated sites. Candidate genes were obtained from a 5000-bp region centered on the region of significant interest.

Reads mapping and variant calling

The clean reads were mapped back to the assembled genome using BWA-MEM 0.7.8-r455 (Li and Durbin, 2009), with modified parameters: the minimum seed length was set to 32, the bandwidth for banded alignment was set to 150 and the penalty for a mismatch was set to 3. PCR duplicates were removed using the rmdup function of SAMTOOLS 0.1.19-44428cd (Li et al., 2009), with default parameters. SNPs were detected using SAMTOOLS 0.1.19-44428cd (Li et al., 2009) and BCFTOOLS 1.3.1 (Narasimhan et al., 2016). The SNPs obtained were further filtered when the mapped reads were <4, the missing data was >0.3 or the minor allele frequencies were <0.05.

Population structure and genetic diversity analysis

A total of 31 samples were used to analyze the population structure of groups of *N. cadamba* with different levels of cadambine content. The cadambine contents of the samples were determined by UPLC-Q-TOF (UPLC1290-6540B Q-TOF; Agilent), as described above (LC-MS/MS). All the filtered SNPs were used to construct a neighbor-joining phylogenetic tree using TREEBEST 1.9.2, with 1000 bootstraps and the other parameters set to default values. The phylogenetic tree was visualized using EVOLVIEW 2 (https://www.evolgenius.info/evolview). In addition, the PCA was performed using the same SNPs set by GCTA 1.24.2 (Yang et al., 2011). The fixation index statistic (F_{ST}) and the nucleotide diversity (π) were calculated using a 40-kb window in 20-kb steps for each population with VCFTOOLS 0.1.14. The selective regions were detected by the top 5% of the low or high $\log_2$ π ratio and the top 5% of the F_{ST} values. The linkage disequilibrium (LD) of each population was estimated by calculating the squared correlation coefficient (r^2) values between any two SNP sites in 500 kb using HAPLOVIEW 4.2 (Barrett et al., 2005).

ACKNOWLEDGEMENTS

We thank Y. Shang for a critical reading of the article. We thank X.S. Hu, L.Z. Gao and Q.G. Liao for helpful discussions. We thank Q. Hu, X. Zhang and S. Xiao for assistance with HPLC-MS/MS experiments. This work was supported by funds from the National Natural Science Foundation of China (grant nos 31470681 and 31970197), the National Key Research and Development Program of China (grant no. 2016YFD0600104), the Natural Science Foundation of Guangdong Province of China (grant no. 2016A030311032), the Guangzhou Science and Technology Program (grant no. 201607020024) and the Foundation of Young Creative Talents in Higher Education of Guangdong Province (grant no. 2017KQNCX017).

AUTHOR CONTRIBUTIONS

CP, XZ and XC designed and managed the project. CP, XZ and XH wrote the article, with input from all authors. KO, QQ, JL and JZ collected the samples. TZ and JG performed the *in vitro* catalytic assay. JY, TZ and XH performed the HPLC-MS/MS experiments. KO, XW, SY and LZ performed the RNA-seq analyses of gene expression. SL and WG performed the FISH analysis. XH, WJ, BL, KO and XZ worked on sequencing and data analysis. CP, XZ, XH, EN, XC and KO revised the article. All the authors have read and approved the final version for publication.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest associated with this work.

DATA AVAILABILITY STATEMENT

All raw and processed sequencing data generated in this study have been submitted to the National Center for Biotechnology Information (NCBI) BioProject database under accession number PRJNA650253. The raw genome sequencing data obtained by Illumina and PacBio platforms have been submitted to the NCBI BioSample database under accession number SAMN15700858. The raw sequencing data of the resequencing data and transcriptome have been submitted to the NCBI BioSample database under accession numbers SAMN15700860 and SAMN15700859, respectively. The genome annotation and assembled genome sequences can be downloaded from https://figshare.com/s/ed20e0e82a4e7474396b.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. A. 17 *k*-mer analysis for estimating the genome size of *Neolamarckia cadamba*. B. Cytological analysis of *N. cadamba* metaphase chromosomes by FISH using *45S* rDNA as the probe. 45S rDNA was labeled with Chroma Tide Alexa Fluor 488 (green signal), and the chromosomes were counterstained with DAPI (blue). Bars = 10 μm.

Figure S2. Estimation of synonymous substitutions per site (K_s) in the *Neolamarckia cadamba* genome.

Figure S3. Characterization of standard compounds by Q-TOF LC-MS/MS and NMR. Tryptamine, expoxystrictosidine, 3α-dihydrocadambine and cadambine used as standards (see "Methods") were analyzed by Q-TOF LC-MS/MS. Cadambine was further tested by NMR to verify its structural formula. NMR spectra were carried out on a Bruker AV 400 spectrometer in CD3OD using tetramethylsilane (TMS, δ = 0) as internal reference. 1H NMR(CD3OD, 400MHz)δ = 7.59 (1H, s), δ = 7.48 (1H, d), δ = 7.34 (1H, d), δ = 7.12 (1H, t), δ = 7.01 (1H, t), δ = 5.85 (1H, d), δ = 4.96 (1H, d), δ = 4.81 (1H, d), δ = 3.90 (1H, m), δ = 3.53 (1H, m), δ = 3.63 (3H, s), δ = 3.56 (1H, m), δ = 3.42 (1H, m), δ = 3.18 (1H, m), δ = 3.03 (1H, m), δ = 2.08 (2H, m), δ = 1.79 (1H, m). In accordance with the results of Handa et al. (1983) and Xu et al. (2011), this compound is characterized as cadambine. A-D MS/MS spectrum of tryptamine (A), cadambine (B), 3α-dihydrocadambine (C) and expoxystrictosidine (D). E. NMR spectrum of cadambine. F. The structural formula of cadambine.

Figure S4. Phylogenetic tree of *STR* genes from six plant species. The genes from *Arabidopsis thaliana* (AT) , *Catharanthus roseus* (CRO), *Coffea canephora* (Cc), Camptotheca acuminate (Cac), *N. cadamba* (evm.model) and *Theobroma cacao* (EOY).

Figure S5. *In vitro* expressed fusion NcSTR1. Left, SDS-PAGE of NcSTR1 expressed in *E. coli* and the strain STR1-2 successfully expressed His-tagged NcSTR1; Middle, *in vitro* expressed fusion NcSTR1 validated by immunoblotting; Right, SDS-PAGE of purified fusion NcSTR1 by His-trap.

Figure S6. Chemical structures of strictosidine, 3α-dihydrocadambine cadambine and the predicted epoxystrictosidine.

Figure S7. The expression profiles of all predicted biosynthetic genes in the shikimate pathway. Bark (B), bud, cambium (C),

young fruit (FR), old leaves (OL), phloem (P), root (R), young leaves (YL), xylem (primary xylem, PX; transitional xylem, TX; secondary xylem, SX), cambium (transitional cambium, TCA; secondary, SCA) and phloem (primary phloem, PPH; transitional phloem, TPH; secondary phloem, SPH) from the first, second and fourth internodes. The second internode of 1-year-old seedling was identified as the transition from primary growth to secondary growth. AnPRT, anthranilate phosphoribosyltransferase; ASA/B, anthranilate synthase; IGPS, indole-3-glycerol phosphate synthase; PRAI, N-(5-phospho-beta-D-ribosyl) anthranilate aldose-ketose-isomerase; SHKA/B, 3-deoxy-7-phosphoheptulonate synthase; SKHC,3-dehydroquinate synthase; SHKD, 3-dehydroquinate dehydratase; SHKE, shikimate dehydrogenase; SHKF, shikimate kinase; SHKG, 3-phosphoshikimate 1-carboxyvinyltransferase; SHKH, chorismate synthase; TDC, tryptophan decarboxylase; TSA-TSB,tryptophan synthase.

Figure S8. The expression profile of all the predicted squalene epoxidase genes (*NcSQEs*). Bark (B), bud, cambium (C), young fruit (FR), old leaves (OL), phloem (P), root (R), young leaves (YL), xylem (primary xylem, PX; transitional xylem, TX; secondary xylem, SX), cambium (transitional cambium, TCA; secondary, SCA) and phloem (primary phloem, PPH; transitional phloem, TPH; secondary phloem, SPH) from the first, second and fourth internodes.

Figure S9. Phylogenetic tree of *NcCYP*s in BGCs and 52 functionally characterized plant CYP family members involved in terpenoid biosynthesis.

Figure S10. Protein alignment of CrLAMT, CcNMTs and their respective *Neolamarckia* paralogs. The aligned sequences were from *Coffea arabica* (Ca, C), *Coffea canephora* (Cc), *Catharanthus roseus* (Cr), *Clarkia breweri* (Cb).

Figure S11. The structure of tandem duplicated *NcDLGT*s.

Figure S12. Heat Map of Expression Data of genes in NcMCL24 and NcMCL25. The sixteen tissues of *N. cadamba* were indicated as those in the Figure S8.

Figure S13. Duplication of NcMYC TFs in tandem order.

Figure S14. Protein alignment of CrSGD, RsSGD, RvSGD, RsRG and their respective *Neolamarckia* and *Ophiorrhiza* paralogs. Accession numbers: CrSGD (AAF28800.1), RsSGD (CAC83098.1), RsRG (AAF03675.1) and RvSGD (AFI71457). Previously reported active-site residues of RsSGD and RsRG were framed or showed with bright colors.

Table S1. Survey statistic results of *Neolamarckia cadamba*.

Table S2. Sequencing data statistics of *Neolamarckia cadamba*.

Table S3. Coverage statistics of the *Neolamarckia cadamba* genome.

Table S4. Assessment of the gene coverage rate using CEGMA and BUSCO.

Table S5. Assessment the genome assembly using RNA-seq.

Table S6. Statistical results for gene function annotation of the *Neolamarckia cadamba* genome.

Table S7. Statistical results for the non-coding RNA of the *Neolamarckia cadamba* genome.

Table S8. Summary of repeat contents in the *Neolamarckia cadamba* genome.

Table S9. Genes used for gene family clustering in each species.

Table S10. Enrichment analysis of duplicated genes from the recent WGD.

Table S11. Genes under positive selection.

Table S12. Predicted biosynthetic genes in five plant species.

Table S13. Short lists of *Neolamarckia cadamba* candidate genes involved in the MVA/MEP, shikimate/indole and iridoid biosynthesis pathways and 35 putative NcSTRs.

Table S14. List of unique genes that underwent a recent WGD event.

Table S15. List of genes in the gene–metabolite association network.

Table S16. The sampling sites of 112 individuals in GWAS.

Table S17. Putative iridoid pathway genes in *Neolamarckia cadamba* associated with loci responsible for MIA production in GWAS.

Table S18. The putative *Neolamarckia cadamba* cytochrome P450s associated with MIA production in GWAS.

Table S19. KEGG analysis of selected genes from population analysis.

Table S20. The 67 biosynthetic gene clusters of *Neolamarckia cadamba* predicted with the plantiSMASH algorithm.

Table S21. Detailed data for the predicted biosynthetic gene clusters and tandem duplicates of *Neolamarckia cadamba*.

REFERENCES

Almeida, A., Dong, A.L., Khakimov, A.B., Bassard, J.E. & Moses, A.T. (2018) A single oxidosqualene cyclase produces the seco-triterpenoid a-onocerin. *Plant Physiology*, **176**(2), 1469–1484.

Barleben, L., Panjikar, S., Ruppert, M., Koepke, J. & Stöckigt, J. (2007) Molecular architecture of strictosidine glucosidase: the gateway to the biosynthesis of the monoterpenoid indole alkaloid family. *The Plant Cell*, **19**, 2886–2897.

Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.

Bedi, Y.S., Bir, S.S. & Gill, B.S. (1981) Cytopalynology of woody taxa of family rubiaceae from North and Central India. *Proceedings of the Indian National Science Academy Part B Biological Sciences*, **47**, 708–715.

Brown, S., Clastre, M., Courdavault, V. & O'Connor, S.E. (2015) De novo production of the plant-derived alkaloid strictosidine in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 3205–3210.

Burton, J.N., Andrew, A., Patwardhan, R.P., Ruolan, Q., Kitzman, J.O. & Jay, S. (2013) Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology*, **31**, 1119.

Carqueijeiro, I., Brown, S., Chung, K., Dang, T.T., Walia, M., Besseau, S. *et al.* (2018a) Two tabersonine 6,7-epoxidases initiate Lochnericine-derived alkaloid biosynthesis in Catharanthus roseus. *Plant Physiology*, **177**, 1473–1486.

Carqueijeiro, I., Dugé de Bernonville, T., Lanoue, A., Dang, T.T., Teijaro, C.N., Paetz, C., Billet, K., Mosquera, A., Oudin, A., Besseau, S. & Papon, N. (2018b) A BAHD acyltransferase catalyzing 19-O-acetylation of tabersonine derivatives in roots of Catharanthus roseus enables combinatorial synthesis of monoterpene indole alkaloids. *The Plant Journal*, **94**, 469–484.

Chae, L., Kim, T., Nilo-Poyanco, R. & Rhee, S.Y. (2014) Genomic signatures of specialized metabolism in plants. *Science*, **344**, 510–513.

Chandel, M., Kumar, M., Sharma, U., Singh, B. & Kaur, S. (2017) Antioxidant, antigenotoxic and cytotoxic activity of Anthocephalus cadamba (Roxb.) Miq. Bark fractions and their phytochemical analysis using UPLC-ESI-QTOF-MS. *Combinatorial Chemistry & High Throughput Screening*, **20**, 760–772.

Chandel, M., Sharma, U., Kumar, N., Singh, B. & Kaur, S. (2012) Antioxidant activity and identification of bioactive compounds from leaves of *Anthocephalus cadamba* by ultra-performance liquid chromatography/electrospray ionization quadrupole time of flight mass spectrometry. *Asian Pacific Journal of Tropical Medicine*, **5**, 977–985.

Chandel, M., Sharma, U., Kumar, N., Singh, B. & Kaur, S. (2014) In vitro studies on the antioxidant/antigenotoxic potential of aqueous fraction

from *Anthocephalus cadamba* bark. In: *Perspectives in Cancer Prevention—Translational Cancer Research*. India: Springer, pp. 61–72.

Chen, Z., Tian, Z., Zhang, Y., Feng, X., Li, Y. & Jiang, H. (2020) Monoterpene indole alkaloids in Uncaria rhynchophlly (Miq.) Jacks chinensis and their chemotaxonomic significance. *Biochemical Systematics and Ecology*, **91**, 104057.

Chin, C.-S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A. *et al.* (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, **13**, 1050.

Colinas, M. & Goossens, A. (2018) Combinatorial transcriptional control of plant specialized metabolism. *Trends in Plant Science*, **23**, 324–336.

de Bie, T., Cristianini, N., Demuth, J.P. & Hahn, M.W. (2006) CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*, **22**, 1269–1271.

de Luca, V.D. & Cutler, A.J. (1987) Subcellular localization of enzymes involved in indole alkaloid biosynthesis in *Catharanthus roseus*. *Plant Physiology*, **85**, 1099–1102.

de Luca, V.D., Salim, V., Thamm, A., Masada, S.A. & Yu, F. (2014) Making iridoids/secoiridoids and monoterpenoid indole alkaloids: progress on pathway elucidation. *Current Opinion in Plant Biology*, **19**, 35–42.

Deng, C., Qin, R., Gao, J., Cao, Y., Li, S., Gao, W. *et al.* (2012) Identification of sex chromosome of spinach by physical mapping of 45s rDNAs by FISH. *Caryologia*, **65**, 322–327.

Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M. *et al.* (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science*, **345**, 1181–1184.

Dwevedi, A., Sharma, K. & Sharma, Y.K. (2014) Cadamba: a miraculous tree having enormous pharmacological implications. *Pharmacognosy Reviews*, **9**, 107–113.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.

Eger, E., Simon, A., Sharma, M., Yang, S., Breukelaar, W.B., Grogan, G. *et al.* (2020) Inverted binding of non-natural substrates in strictosidine synthase leads to a switch of stereochemical outcome in enzyme-catalyzed Pictet-Spengler reactions. *Journal of the American Chemical Society*, **142**, 792–800.

Handa, S.S., Borris, R.P., Cordell, G.A. & Phillipson, J.D. (1983) NMR spectral analysis of cadambine from *Anthocephalus chinensis*. *Journal of Natural Products*, **46**, 325–330.

Hu, L., Xu, Z., Wang, M., Fan, R., Yuan, D., Wu, B. *et al.* (2019) The chromosome-scale reference genome of black pepper provides insight into piperine biosynthesis. *Nature Communications*, **10**, 4702.

Jacobowitz, J.R. & Weng, J.-K. (2020) Exploring uncharted territories of plant specialized metabolism in the postgenomic era. *Annual Review of Plant Biology*, **71**, 631–658.

Kai, G., Wu, C., Gen, L., Zhang, L., Cui, L. & Ni, X. (2015) Biosynthesis and biotechnological production of anti-cancer drug Camptothecin. *Phytochemistry Reviews*, **14**, 525–539.

Kareti, S.R. & Subash, P. (2020) In silico exploration of anti-Alzheimer's compounds present in methanolic extract of Neolamarckia cadamba bark using GC–MS/MS. *Arabian Journal of Chemistry*, **13**, 6246–6255.

Kautsar, S.A., Suarez Duran, H.G., Blin, K., Osbourn, A. & Medema, M.H. (2017) plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Research*, **45**, W55–W63.

Kellner, F., Kim, J., Clavijo, B.J., Hamilton, J.P., Childs, K.L., Vaillancourt, B. *et al.* (2015) Genome-guided investigation of plant natural product biosynthesis. *The Plant Journal*, **82**, 680–692.

Ketudat Cairns, J.R. & Esen, A. (2010) β-Glucosidases. *Cellular and Molecular Life Sciences*, **67**, 3389–3405.

Kim, D., Langmead, B. & Salzberg, S.L. (2015a) HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, **12**, 357–360.

Kim, D., Langmead, B. & Salzberg, S.L. (2015b) HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, **12**, 357–360.

Kim, D., Paggi, J.M., Park, C., Bennett, C. & Salzberg, S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, **37**, 907–915.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. & Salzberg, S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, **14**(4), 1–13.

Leonelli, L., Brooks, M. & Niyogi, K.K. (2017) Engineering the lutein epoxide cycle into *Arabidopsis thaliana*. *Proceedings of the National Academy of Ences*, **114**, 201704373.

Li, H. & Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li, J., Zhang, D., Ouyang, K. & Chen, X. (2018) The complete chloroplast genome of the miracle tree *Neolamarckia cadamba* and its comparison in Rubiaceae family. *Biotechnology & Biotechnological Equipment*, **32**, 1087–1097.

Li, J., Zhang, D., Ouyang, K. & Chen, X. (2019) High frequency plant regeneration from leaf culture of Neolamarckia cadamba. *Plant biotechnology (Tokyo, Japan)*, **36**(1), 13–19.

Li, L., Stoeckert, C.J. & Roos, D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, **13**, 2178–2189.

Li, M., Tian, S., Jin, L., Zhou, G., Li, Y., Zhang, Y. *et al.* (2013) Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. *Nature Genetics*, **45**(12), 1431–1438.

Li, Y., Leveau, A., Zhao, Q., Feng, Q. & Osbourn, A. (2021) Subtelomeric assembly of a multi-gene pathway for antimicrobial defense compounds in cereals. *Nature Communications*, **12**, 2563.

Ma, D. & Constabel, C.P. (2019) MYB repressors as regulators of phenylpropanoid metabolism in plants. *Trends in Plant Science*, **24**, 275–289.

Ma, X., Koepke, J., Panjikar, S., Fritzsch, G. & Stockigt, J. (2005) Crystal structure of vinorine synthase, the first representative of the BAHD superfamily. *Journal of Biological Chemistry*, **280**, 13576–13583.

Ma, X., Panjikar, S., Koepke, J., Loris, E. & Stöckigt, J. (2006) The structure of Rauvolfia serpentina strictosidine synthase is a novel six-bladed β-propeller fold in plant proteins. *The Plant Cell*, **18**, 907–920.

Mccarthy, A.A. & Mccarthy, J.G. (2007) The structure of two N-methyltransferasesfrom the caffeine biosynthetic pathway. *Plant Physiology*, **144**(2), 879–889.

Mehra, P. & Bawa, K. (1969) Chromosomal evolution in tropical hardwoods. *Evolution*, 466–481.

Meijer, A.H., Cardoso, M., Voskuilen, J.T., Waal, A.D., Verpoorte, R. & Hoge, J. (1993) Isolation and characterization of a cDNA clone from Catharanthus roseus encoding NADPH:cytochrome P-450 reductase, an enzyme essential for reactions catalysed by cytochrome P-450 mono-oxygenases in plants. *The Plant Journal*, **4**(1), 47–60.

Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C. & Durbin, R. (2016) BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics*, **32**, 1749–1751.

Nützmann, H.W., Huang, A. & Osbourn, A. (2016) Plant metabolic clusters–from genetics to genomics. *New Phytologist*, **211**, 771–789.

Pandey, A. & Negi, P.S. (2016) Traditional uses, phytochemistry and pharmacological properties of Neolamarckia cadamba: A review. *Journal of Ethnopharmacology*, **181**, 118–135.

Parra, G., Bradnam, K. & Korf, I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061–1067.

Paterson, A.H., Bowers, J.E. & Chapman, B.A. (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 9903–9908.

Pressnitz, D., Fischereder, E.M., Pletz, J., Kofler, C., Hammerer, L., Hiebler, K. *et al.* (2018) Asymmetric synthesis of (R)-1-Alkyl-Substituted Tetrahydro-ß-carbolines catalyzed by strictosidine synthases. *Angewandte Chemie*, **130**, 10843–10847.

Puttick, Mark N. (2019) MCMCtreeR: functions to prepare MCMCtree analyses and visualize posterior ages on trees. *Bioinformatics*, **35**, 5321–5322.

Qu, Y., Easson, M.E., Simionescu, R., Hajicek, J., Thamm, A.M., Salim, V. *et al.* (2018) Solution of the multistep pathway for assembly of corynanthean, strychnos, iboga, and aspidosperma monoterpenoid indole alkaloids from 19E-geissoschizine. *Proceedings of the National Academy of Sciences of the United States of America*, **115**, 3180–3185.

Rai, A., Hirakawa, H., Nakabayashi, R., Kikuchi, S., Hayashi, K., Rai, M. *et al.* (2021) Chromosome-level genome assembly of Ophiorrhiza pumila

reveals the evolution of camptothecin biosynthesis. *Nature Communications*, **12**, 1–19.

Razafimandimbison, S.G. (2002) A systematic revision of Breonia (Rubiaceae-Naucleeae). *Annals of the Missouri Botanical Garden*, **89**, 1–37.

Robbrecht, E. & Manen, J.F. (2006) The major evolutionary lineages of the coffee family (Rubiaceae, Angiosperms). Combined analysis (nDNA and cpDNA) to infer the position of Coptosapelta and Luculia, and Supertree construction based on rbcL, rps16, trnL-trnF and atpB-rbcL Dat. A new classification in two subfamilies, Cinchonoideae and Rubioideae. *Systematics and Geography of Plants*, **76**, 85–145.

Rosseleena, R.E., Motoaki, C., Miki, K., Takashi, A., Yoshimi, O., Nobutaka, M. *et al.* (2016) An MYB transcription factor regulating specialized metabolisms in Ophiorrhiza pumila. *Plant Biotechnology*, **33**, 1–9.

Sabir, J.S., Jansen, R.K., Arasappan, D., Calderon, V., Noutahi, E., Zheng, C. *et al.* (2016) The nuclear genome of *Rhazya stricta* and the evolution of alkaloid diversity in a medically relevant clade of Apocynaceae. *Scientific Reports*, **6**, 1–10.

Sadre, R., Magallanes-Lundback, M., Pradhan, S., Salim, V., Mesberg, A., Jones, A.D. *et al.* (2016) Metabolite diversity in alkaloid biosynthesis: a multilane (diastereomer) highway for camptothecin synthesis in Camptotheca acuminata. *The Plant Cell*, **28**, 1926–1944.

Salim, V., Yu, F., Altarejos, J. & de Luca, V. (2013) Virus-induced gene silencing identifies *Catharanthus roseus* 7-deoxyloganic acid-7-hydroxylase, a step in iridoid and monoterpene indole alkaloid biosynthesis. *The Plant Journal*, **76**, 754–765.

Santos, C.L., Angolini, C.F., Neves, K.O., Costa, E.V., de Souza, A.D., Pinheiro, M.L., Koolen, H.H. & da Silva, F.M. (2020) Molecular networking-based dereplication of strictosidine-derived monoterpene indole alkaloids from the curare ingredient Strychnos peckii. *Rapid Communications in Mass Spectrometry*, **34**, e8683.

Schläpfer, P., Zhang, P., Wang, C., Kim, T., Banf, M., Chae, L., Dreher, K., Chavali, A.K., Nilo-Poyanco, R., Bernard, T. & Kahn, D. (2017) Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant Physiology*, **173**, 2041–2059.

Scott, M.G. & Madden, T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, **32**, 20–25.

Shang, Y. & Huang, S. (2020) Engineering plant cytochrome P450s for enhanced synthesis of natural products: past achievements and future perspectives. *Plant Communications*, **1**, 100012.

Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.

Singh, S.K., Patra, B., Paul, P., Liu, Y., Pattanaik, S. & Yuan, L. (2021) BHLH IRIDOID SYNTHESIS 3 is a member of a bHLH gene cluster regulating terpenoid indole alkaloid biosynthesis in Catharanthus roseus. *Plant Direct*, **5**, e00305.

Stamatakis, Alexandros (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Takayama, H., Tsutsumi, S.-I., Kitajima, M., Santiarworn, D., Liawruangrath, B. & Aimi, N. (2003) Gluco-indole alkaloids from Nauclea cadamba in thailand and transformation of 3α-dihydrocadambine into the indolopyridine alkaloid, 16-carbomethoxynaufoline. *Chemical and Pharmaceutical Bulletin*, **51**, 232–233.

Tatsis, E.C., Carqueijeiro, I., de Bernonville, T.D., Franke, J., Dang, T.T., Oudin, A., Lanoue, A., Lafontaine, F., Stavrinides, A.K., Clastre, M. & Courdavault, V. (2017) A three enzyme system to generate the Strychnos alkaloid scaffold from a central biosynthetic intermediate. *Nature Communications*, **8**, 316.

Tran, H.T., Ramaraj, T., Furtado, A., Lee, L.S. & Henry, R.J. (2018) Use of a draft genome of coffee (Coffea arabica) to identify SNP s associated with caffeine content. *Plant Biotechnology Journal*, **16**, 1756–1766.

Trapnell, C., Pachter, L. & Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R. *et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, **7**, 562–578.

Urlacher, V.B. & Girhard, M. (2019) Cytochrome P450 monooxygenases in biotechnology and synthetic biology. *Trends in Biotechnology*, **37**, 882–897.

van Moerkercke, A., Steensma, P., Gariboldi, I., Espoz, J., Purnama, P.C., Schweizer, F. *et al.* (2016) The basic helix-loop-helix transcription factor BIS2 is essential for monoterpenoid indole alkaloid production in the medicinal plant Catharanthus roseus. *The Plant Journal*, **88**, 3–12.

van Moerkercke, A., Steensma, P., Schweizer, F., Pollier, J., Gariboldi, I., Payne, R. *et al.* (2015) The bHLH transcription factor BIS1 controls the iridoid branch of the monoterpenoid indole alkaloid pathway in Catharanthus roseus. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 8130–8135.

Wang, C., Wu, C., Wang, Y., Xie, C., Shi, M., Nile, S. *et al.* (2019) Transcription factor OpWRKY3 is involved in the development and biosynthesis of camptothecin and its precursors in *Ophiorrhiza pumila* hairy roots. *International Journal of Molecular Sciences*, **20**, 3996.

Wang, D.-P., Wan, H.-L., Zhang, S. & Yu, J. (2009) Gamma-MYN: a new algorithm for estimating Ka and Ks with consideration of variable substitution rates. *Biology Direct*, **4**, 20.

Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X. *et al.* (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research*, **40**, e49.

Weir, B.S. & Cockerham, C.C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

Wu, X.D., Wang, L., He, J., Li, X.Y., Dong, L.B., Gong, X. *et al.* (2013) Two new indole alkaloids from *Emmenopterys henryi*. *Helvetica Chimica Acta*, **96**, 2207–2213.

Wu, S., Yang, M. & Xiao, Y. (2018) Synthetic biology studies of monoterpene indole alkaloids. *Chinese Journal of Organic Chemistry*, **38**, 2243–2258.

Xia, L., Ruppert, M., Wang, M., Panjikar, S., Lin, H., Rajendran, C. *et al.* (2012) Structures of alkaloid biosynthetic glucosidases decode substrate specificity. *ACS Chemical Biology*, **7**, 226–234.

Yamazaki, Y., Sudo, H., Yamazaki, M., Aimi, N. & Saito, K. (2003) Camptothecin biosynthetic genes in hairy roots of Ophiorrhiza pumila: cloning, characterization and differential expression in tissues and by stress compounds. *Plant and Cell Physiology*, **44**, 395–403.

Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. (2011) GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, **88**, 76–82.

Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**, 1586–1591.

Yang, Y., Li, W., Pang, J., Jiang, L., Qu, X., Pu, X. et al. (2019) Bifunctional cytochrome P450 enzymes involved in camptothecin biosynthesis. *ACS Chemical Biology*, **14**(6), 1091–1096.

Yuan, H.-L., Zhao, Y.-L., Qin, X.-J., Liu, Y.-P., Yu, H.-F., Zhu, P.-F. *et al.* (2020) Anti-inflammatory and analgesic activities of Neolamarckia cadamba and its bioactive monoterpenoid indole alkaloids. *Journal of Ethnopharmacology*, **260**, 113103.

Zhang, H., Hedhili, S., Montiel, G., Zhang, Y., Chatel, G., Pré, M. *et al.* (2011) The basic helix-loop-helix transcription factor CrMYC2 controls the jasmonate-responsive expression of the ORCA genes that regulate alkaloid biosynthesis in Catharanthus roseus. *The Plant Journal*, **67**, 61–71.

Zheng, X., Li, P. & Lu, X. (2019) Research advances in cytochrome P450-catalysed pharmaceutical terpenoid biosynthesis in plants. *Journal of Experimental Botany*, **70**, 4619–4630.

Zhou, X. & Stephens, M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, **44**, 821–824.

Zhou, X. & Stephens, M. (2014) Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, **11**, 407–409.