DR XIAOLAN ZHAO (Orcid ID : 0000-0001-7130-9388)

MISS XIAODI HU (Orcid ID : 0000-0001-9704-2474)

Article type : Original Article

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

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

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

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

[3]College of Life Sciences, Henan Normal University, Xinxiang 453007 China

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

5 Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, Michigan 48109, USA

[6] These authors contributed equally to this article.

Correspondence: Changcao Peng (ccpeng@scau.edu.cn), Xiaoyang Chen (xychen@scau.edu.cn)

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

Significance statement: *Neolamarckia cadamba*, *Coffea canephora* and *Ophiorrhiza pumila* are three closely-related species in the Rubiaceae family which produce quite different major specialized metabolites including cadambine, caffeine (a purine alkaloid) and camptothecin (anti-cancer MIAs). This chromosome-level assembly of *N. cadamba* genome and cadambine synthetic pathway analysis can not only facilitate the development of tools for enhancing bioactive productivity by metabolic engineering but also accelerate the understanding of the evolutionary history of specific metabolic pathways.

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 accumulated in the bark and leaves, exhibit antimalarial, antiproliferative, antioxidant, anticancer and anti-inflammatory activity. 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 *C. canephora* revealed that *N. cadamba* underwent a relatively recent whole-genome duplication (WGD) event after diverging from *C. canephora* which contributed to the evolution of MIA biosynthetic pathway. We determined the key intermediates of 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* (De Luca *et al.*, 2014, Kai *et al.*, 2015, Sadre *et al.*, 2016, Tran *et al.*, 2018, Rai *et al.*, 2021). *N. cadamba* was called a ‘miracle tree’ by World Forestry Congress (WFC) in

1972 because 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, anaemia, stomatitis, leprosy, cancer and a variety of infectious diseases in Southeast Asia (Pandey and Negi, 2016). Although, there is lack of in-depth studies to clarify the active metabolites responsible for various pharmacological activities attributed to *N. cadamba*, recent studies pinpoint 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 activity (Chandel *et al.*, 2014, Dwevedi *et al.*, 2014, Chandel *et al.*, 2017, Yuan *et al.*, 2020).

Since its original discovery in *N. cadamba (syn. Anthocephalus chinensis)*, cadambine was also found to accumulate in *Uncaria spp., Haldina cordifolia* and *Emmenopterys henryi* species in the Rubiaceae family (Handa *et al.*, 1983, Wu *et al.*, 2013, Wang *et al.*, 2019, Chen *et al.*, 2020). The structure of cadambine was firstly 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 originates from the common precursor 3-a(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, determination of the key intermediate strictosidine and characterization of functional strictosidine synthase (STR) is the prerequisite for elucidation of cadambine biosynthetic pathway in *N. cadamba.*

Here we report a chromosome-level genome assembly of *N. cadamba* (2n = 44 chromosomes) obtained by a combination of Illumina and PacBio data platform. We used a 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 *C. canephora* and twelve other available plant genomes to investigate whole-genome

duplication (WGD) events and 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 will likely 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 ~ 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 (123.25X) generated from PacBio Sequel platform, plus another 52.30 Gb reads (69.33X) from Illumina platform, 184.45 Gb 10X Genomics data (244.52X) and 89.59Gb Hi-C data (118.81X) (**Table S2**). The Falcon software package (Chin *et al.*, 2016) was used for initial assembly of the PacBio reads, then polished error-corrected with both PacBio and Illumina reads. 10X Genomics data was used to anchor contigs into scaffolds. Finally, with the 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 (**Table 1 &Figure 1a**). The total length of the assembly was 744.5 Mb, which represents 98.7% of the estimated genome size.

To evaluate the assembly quality, 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, Kim *et al.*, 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 indicated 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 3,489 base pairs and an average of 4.7 exons per gene (**Table 1**). Approximately 96.4% of the genes with shared homology to NR, Swiss-Prot, KEGG and InterPro databases known genes 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, 1,642 ribosomal RNA (rRNA, 18S, 28S, 5.8S and 5S) genes, 2,701 small nuclear RNA (snRNA) genes and 1,053 microRNA (miRNA) genes.

Repeat annotation revealed that 52.9% (394.1 Mb) of the assembled *N. cadamba* genome comprises of transposable elements (TEs) (**Figure1b**&**Table S8**). Retrotransposons were found to be the dominant class of repeat elements (48.2%), while 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 shared across 14 species. Phylogenetic analysis revealed that *N. cadamba* diverged from *Solanum tuberosum* (Solanaceae), *O. pumila* (Rubiaceae) and *C. canephora* (Rubiaceae) at

around 80.4 MYA, 41.7 MYA and 31.0 MYA, respectively (**Figure 1c**). Compared with *C. 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 pathways were significantly enriched (FDR cut-off < 0.05) which 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 GH1 β- glucosidase (BGLU) family, which are not only involved in starch and sucrose metabolism, but also play a role in the biosynthesis of phenylpropanoid and secondary metabolite (Ketudat and 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 distance = ~0.17 and 4DTv distance = ~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) occurred ~70 MYA in core eudicots (Paterson *et al.*, 2004). The 4DTv distribution also recovered the WGT-γ in *C. 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 *C. canephora* (**Figure 1d**), which was further supported by the distribution of the synonymous substitution rate (*Ks*) (**Figure S2**). Intergenomic co-linearity analysis demonstrated a 2:1 syntenic relationship between *N. cadamba* and *C. canephora,* and 186 syntenic blocks were identified in *N. cadamba* by comparing to the *C. 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 possibly under positive selection (PSGs, p-value<0.01, FDR < 0.05) using the branch-site

model of the PAML software (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 firstly confirmed the structure of cadambine by Q-TOF LC-MS/MS (**Figure 2a; Figure S3b**) (Chandel *et al.*, 2012, Chandel *et al.*, 2017) and NMR spectra (**Figure S3e, S3f) (Handa *et al.*, 1983)**. The standard tryptamine (**Figure 2a; Figure S3a**), 3α-dihydrocadambine (**Figure 2b; Figure S3c**) **(Takayama *et al.*, 2003)** and epoxystrictosidine (**Fig 2b; Figure S3d**) were also analyzed by high resolution mass spectrometry. Secondly, the extracts from the leaves and barks of *N. cadamba* were analyzed, this analysis revealed that *N. cadamba* accumulated tryptamine (**Figure 2c, 2f**) in the leaves, and strictosidine **(Figure 2c, 2g)**, epoxystrictosidine (**Figure 2d, 2i**), 3α-dihydrocadambine (**Figure 2d, 2h**) and cadambine (**Figure 2e, 2j**) in the bark. There were two compounds resolved with the same of m/z values (547.2285) (**Figure 2h, 2i**), one of which was matched with the peak at 12.30 min as standard 3α-dihydrocadambine ($C_{27}H_{34}N_2O_{10}$), another with the peak at 13.57 min as standard epoxystrictosidine ($C_{27}H_{34}N_2O_{10}$) **(Figure 2d)**. Notably, the key intermediate strictosidine was detected. MS/MS spectrum of the ion at m/z 531 (**Figure 2g**) displayed the same fragmentation pattern such as 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 STR family expanded in two Rubiaceae species (35 copies and 28 copies in *N. cadamba* and coffee, respectively) compared to *C. roseus*, *T. cacao* and *A. thaliana*. Of the 35 putative *NcSTRs* **(Figure 3b)**, seven *NcSTRs* were grouped with a conserved clade

containing characterized strictosidine synthase gene *CRO T006099* (Kellner *et al.*, 2015) (**Figure S4)** ,whereas only five *NcSTRs* (named as *NcSTR1, 13,14, 22, 27*) share over 45% aa identities to CrSTR. **S**equence 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, 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 *N. cadamba* genome **(Figure 6a)**, and the predicted *NcTDC2* protein shares 71 and 84% sequence identities with the functional tryptophan decarboxylases of *C. roseus* (De Luca and Cutler, 1987) and *O. pumila* (Yamazaki *et al.*, 2003a), respectively, thus, we firstly examined the catalytic activities of the expressed NcSTR1 enzyme.

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**), a reaction product of strictosidinewas produced and detected by HPLC-MS/MS. Two ion pairs, 531.10 ->352.20 (blue curve) and 531.10 ->514.25 (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 hours reaction comparing with 24 hours (**Figure 3c**). The control without NcSTR1 did not have the relative peak. These results indicated 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 a same upstream pathway to the formation of strictosidine as that of *C. roseus* (**Figure 4a**). The gene lists involved in MVA/ MEP and shikimate/indole

pathway 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 S 13**). The predicted genes 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 the orthologues with more than 65% aa identity to functional genes from C. *roseus* and *O. pumila* (Salim *et al.*, 2013, Brown *et al.*, 2015, Kai *et al.*, 2015, Kellner *et al.*, 2015). 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 *C. canephora* (Ks<1 in **Table S 14**), indicating that the recent WGD event were 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 the strictosidine followed by 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 (**Figure 4a and Figure S6).** In consistence 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 the possible organizers in the cadambine synthetic pathway (Ma *et al.*, 2005, Xia *et al.*, 2012, Leonelli *et al.*, 2017, Almeida *et al.*, 2018, Carqueijeiro *et al.*, 2018a, Carqueijeiro *et al.*, 2018b, Qu *et al.*, 2018, Zheng *et al.*, 2019, Shang and Huang, 2020, Tatsis *et al.*, 2017). In order to identify more biosynthetic gene candidates, we further conducted RNA-seq analysis, GWAS and 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* which 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* (**Figure 4b** and **Figure S7**), indicating that the synthesis of secoiridoid was coordinated with indole precursors needed for cadambine production. In addition, NcSTR1/ 14/ 19/ 29/ 35 and NcSQE1/6 exhibited co-expression pattern with that of *NcLAMT1/2* and *NcTDC2* (**Figure 3b, 4b** and **Figure 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 (FPKM) and the metabolites contents (μg) per gram fresh weight of five tissues were determined using a liquid chromatography–mass spectrometry-based method. We calculated the Pearson correlations among the contents of strictosidine, tryptamine, cadambine and gene transcript levels, respectively. The correlation analysis (coefficient >0.9 as the cutoff) showed that the expression level of *NcSTR1* was strongly associated with strictosidine content, and that *NcSTR14* (evm.model.Contig341.115), *NcSLS3* (evm.model.Contig60.399) with cadambine content (**Table S15**). Two *NcCYPs* (evm.model.Contig208.74, evm.model.Contig 708.57) and *NcOSC1* (evm.model.Contig207.232*)* were also identified to be strongly associated with cadambine content in the network. Both NcCYPs were assigned into 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 twelve transcription factors (TFs) were highly correlated with cadambine accumulation pattern. Of these, *evm.model.Contig 28.407* putatively encoded a basic helix–loop–helix (bHLH) TF with 50% and 45% aa identity to 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 cadambine synthetic pathway, we conducted a genome-wide association study (GWAS) using 112 individuals collected from 27 populations distributed widely in Southeast Asia (**Table S16**). We identified 5,786,667 high-quality SNPs based on the reference genome. To establish 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 *NcCYPs* putatively involved in the epoxidation and the hydrolysis of epoxide (Meijer *et al.*, 1993, Zheng *et al.*, 2019) 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 level. 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 and 5b**). We observed similar results in phylogenetic tree and PCA analyses.

We also analyzed the linkage disequilibrium (LD) throughout the whole genome. LD (indicated by r2) 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 low cadambine

population and slowest in high (**Figure 5c**). Then, we explored the genomic regions with high divergence to try to identify genes possibly involved in the 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 F_{ST} values between ecotype groups were identified as selective sweeps.

We calculated the fixation index (F_{ST}) between subgroups, and the genomic regions with fixation index (F_{ST}) values (Weir and Cockerham, 1984) in the top 5% were considered highly differentiated. We estimated the population diversification parameters, π and θw ，and found that the overall nucleotide diversity in species with high cadambine level was higher than that in species with low- or medium-cadambine level (high- cadambine level species, π=0.0023440 and θw=0.0019451; low- cadambine level species, π= 0.0023134 and θw= 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** and **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 (Nützmann *et al.*, 2016, Schläpfer *et al.*, 2017, Jacobowitz and Weng, 2020). We systematically mined the *N. cadamba* genome sequence to identify all the BGCs with 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* within physical sizes less than 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 *CYP71Ds* subfamily members, a *CYP76* member and four *TPSs* **(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 *R. stricta* (Sabir *et al.*, 2016) , *G. sempervirens* and *O. pumila* (Rai *et al.*, 2021) . All the six *NcCYPs* in this cluster were derived from the recent WGD event (Ks<1) , and the closest match of the five NcCYP71D paralogs is *C. 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, 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 Mb-25.13 Mb) and pseudochromosome 22 (3.78 Mb- 5.73 Mb) were identified, in which 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 to CrDL7H) is located c. 600 kb from NcLAMT1, and two other *NcCYP72As* were closely adjacent to the predicted NcDL7H. Interestingly, the two cadambine biosynthetic candidates (evm.model.Contig208.74, evm.model.Contig 708.57) identified in the gene-to-metabolite network were also distributed in NcSMASH67 and NcMCL23, respectively, implying a selection pressure favoring 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 NcCYP81s (**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 *C. canephora* further revealed that *N. cadamba* had underwent a WGD. This is further supported by the low median Ks value (~ 0.41) of gene paralogs for most genes in NcMCL15 after excluding the two key genes NcIS1 and NcIS2. NcSQE1 is located c. 300 kb from NcIS1 in cluster NcMCL15, and the co-expression profiles of NcSQE1/6, NcLAMT1/2, NcTDC2 and NcSTR1/14 suggests that NcSQE1/6 might be involved in the epoxidation of the strictosidine.

Another predicted cluster (NcSMASH17) in *N. cadamba* is located on pseudochromosome 4 and consists of NcSQE4, a MYB TFs 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-to-metabolite network (**Table S15**). OpMYB1 is an R2R3- MYB repressor that acted as a negative regulator of MIA production, and its overexpression in hairy roots of *O. pumila* resulted in reduced production of camptothecin and reduced expression of OpTDC (Rosseleena *et al.*, 2016, Ma and Constabel, 2019).

Previous studies demonstrated that specialized metabolic genes were more significantly enriched in local (tandem) duplication events as compared with whole-genome duplication 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 (**Figure 6a,6b** and **Figure S10**).

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 *C. 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, Van Moerkercke *et al.*, 2016, Singh *et al.*, 2021). In the gene-to-metabolite network, we characterized one homologue to *CrBIS1*, evm.model.Contig 28.407. Analysis of the genome sequence

of *N. cadamba* showed that four additional bHLH TF genes clustered with evm.model.Contig 28.407 in a tandem duplication within a 130-kb region on pseudochromosome 4 (NcMCL24, **Figure 6d**). Synteny analysis further detected another bHLH TFs tandem duplicates (NcMCL25) with close homology and gene localization to those of NcMCL24 on pseudochromosome 8, in which consists of the 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 suggesting 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 characterized 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) to CrMYC2 (AF283507), the major activator of MeJA-responsive ORCA2/3, which in turn regulated a subset of alkaloid biosynthesis genes in *C. 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. *N. cadamba* was characterized as a tetraploid species in those relatively old references (Mehra and Bawa, 1969, Bedi *et al.*, 1981). As there were confusing names for *N. cadamba* (Pandey and Negi, 2016), we first identified the sequenced *N. cadamba* is diploid with 2n = 2x =44 chromosomes by chromosome measurements and a pair of 45s rDNA probe signals using fluorescent in situ hybridization (FISH) (**Figure S1B**). This provided a solid base for establishing a high-quality of *N. cadamba* Hi-C libraries.

To date, three species in the genus *Coffea* (*C. canephora*, *C. arabica*, *C. eugenioides*), *O. pumila,* and *Gardenia jasminoides* in the family Rubiaceae have been sequenced. *N. 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 *N. macrophyllus,* and the latter is endemic to Sulawesi, Indonesia (Li *et al.*, 2018). This genus is featured by the densely globe-shaped flower clusters (the origin of its Chinese name "TuanHua") with 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 accumulated high content (~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 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 *N. cadamba*, *C. canephora* and *O. pumila* will shed light on the evolutionary history of specific metabolic pathways as they produce quite different major specialised 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, we found that none candidate homologous (aa identity cutoff >55%) corresponding to the late seco-iridoid pathway genes, *NcLAMT1*/*NcLAMT2* and *NcSLS1-4*, were identified in *C. canephora* genome. In contrast to coffea, the Ophiorrhiza candidate homologous OpLAMT (Opuchr05_g0056110) and OpSLSs (Opuchr02_g0012990, Opuchr02_g0013060, Opuchr02_g0013090 and Opuchr02_g0017930) share 84% aa identity with NcLAMT1/NcLAMT2 and 67-85% aa identities with NcSLS1-4, respectively. *N. cadamba* diverged ~ 41.7 Mya from *O. pumila* and ~ 31.0 Mya from *C. canephora* (**Figure 1a**), we therefore propose that the LAMT gene evolution involved a deletion following a duplication process, and that the duplication of NcLAMT derived from a recent WGD event was key for evolution of cadambine biosynthesies in *N. cadamba* as there is only one copy of LAMT gene in both well-known MIA-producing species, such as *C. 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 *C. 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, Ile 266, Ser 316 and Tyr 356 for CcXMT1 (Mccarthy A A and Mccarthy J G, 2007) were no longer conserved in the deduced amino acid sequences of the Neolamarckia homologous, 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, EC 3.2.1.105) and raucaffricine O-β-D-glucosidase (RG, EC 3.2.1.125), act as major components in the MIA biosynthesis pathway. In *C. 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 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 displayed that all the Neolamarckia *and* Ophiorrhiza homologus 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 not conservative at four critical active site residues (Thr189, His193, Tyr200, 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

utilization of germplasm. Moreover, new genetic variation is needed to use to increase the cadambine content. In our study, we reported the genome variation mapping of 112 accessions and we 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 of selective sweeps associated with agriculturally important markers will be helpful for molecular breeding in *N. cadamba*.

Methods

Chromosome preparation and FISH analysis

For improved characterization of the chromosomes, 45S rDNA was labeled with Chroma Tide Alexa Fluor 488–5-dUTP (Invitrogen) for FISH and theitotic metaphase spreads were prepared from meristem root tip cells of *N. cadamba* following the procedures described in the reference (Deng *et al.*, 2012) with minor modifications.

Plant Materials, DNA Extraction and Genome Sequencing

The young tender leaves of *N. cadamba* collected from a 7-years-old individual plant in South China Agricultural University (Guangzhou, China) were used for DNA extraction. Paired-end (PE) library (insert size 350 bp) was constructed and sequenced on Illumina Xten platform.

SMRT Bell libraries with an insert size of 20 kb were also constructed and sequenced on PacBio Sequel platform. For 10 X genomic sequencing, DNA sample preparation, indexing, and barcoding were done using the GemCode Instrument (10 X Genomics).

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, it 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 to the initial assembly using BWA and scaffolding approach was performed by FragScaff. After that, we used Pbjelly software 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 by the LACHESIS pipeline.

To evaluate the completeness and quality of the assembly, the high-quality reads from short insert size 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 assessment the completeness of the genome assembly.

Repetitive elements and genes annotation

The approach combined the *de novo* and homology was used to search transposable elements (TEs) in the *N. cadamba* genome. In *de novo* approach, RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html), RepeatScout, LTR-Finder and Tandem Repeats Finder (TRF) sofrwares were used. In homology prediction, RepeatMasker (version 3.3.0) and RepeatProteinMask software were used to against Repbase TE library (http://www.girinst.org/repbase) and 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 proteins sequences of *Arabidopsis thaliana*, *Populus trichocarpa*, *Solanum tuberosum* and *Coffea canephora* were

downloaded from Ensemble (http://plants.ensembl.org/index.html) and NCBI (https://www.ncbi.nlm.nih.gov/) and then aligned to *N. cadamba* genome assembly using tblastn (E-value $1e^{-5}$).

Ab initio gene prediction: Augustus (version 2.5.5), Genscan (version 1.0), GlimmerHMM (version 3.0.1), Geneid and SNAP were used to predict coding regions in the repeat-masked genome.

Transcriptome-assisted gene prediction: Tophat (version 2.0.8)(Trapnell *et al.*, 2009, Kim *et al.*, 2013) was used to map clean RNA-seq reads to *N. cadamba* genome and Cufflinks (version 2.1.1)(Trapnell *et al.*, 2012) (http://cufflinks.cbcb.umd.edu/) and then used to assemble the transcripts into gene models (Cufflinks-set).

Gene model evidence from Homo-set, ab initio and Cufflinks-set programs 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. The InterPro and Pfam databases search were performed using InterproScan (V4.8) and HMMER (V3.1), respectively. For the other databases, BLAST searches were performed with an E value cutoff of 1e−05.

Species phylogenetic analysis

OrthoMCL(http://orthomcl.org/orthomcl/) was used to cluster paralogous and orthologous among 14 species (*N. cadamba*, *Arabidopsis thaliana*, *C. canephora*, *O. pumila* , *Eucalyptus grandis*, *Elaeis guineensis*, *Malus domestic*, *Glycyrrhiza uralensis*, *Theobroma cacao*, *Oryza sativa*, *S. tuberosum*, *Populus trichocarpa*, *Actinidia chinensis* and *Vitis vinifera*). We obtained the similarity relations between all species protein sequences through all-vs-all blastp with the e-value 1e-5. MUSCLE(Edgar, 2004) (http://www.drive5.com/muscle/) was used to align all 402

single-copy gene protein sequences and make a super alignment matrix. Then, RAxML(Stamatakis, 2014) (http://sco.h-its.org/exelixis/web/software/raxml/index.html) was used to construct the 14 species phylogenetic tree with maximum likelihood method. Finally, the MCMCtree program(Puttick, 2019) (http://abacus.gene.ucl.ac.uk/software/paml.html) was applied to infer the divergence time based on the phylogenetic tree constructed. CAFÉ 2.2(De Bie *et al.*, 2006) was used to determin expansion and contractions of orthologous gene families.

Whole genome duplication (WGD)

To identity syntenic blocks, the protein sequences from *C. canephora*, *S. tuberosum* and *N. cadamba* were searched against themselves using blastp (E<1e-5)(Scott and Madden, 2004). Syntenic blocks were determined by MCscanX (Wang *et al.*, 2012) with the parameter at least five genes per block. Then we calculated the 4DTv (fourfold degenerate synonymous sites of the third codons) for syntenic segments from the concatenated alignments which constructed by fourfold degenerate sites of all gene pairs found in each segment and plotted the distribution of the 4DTv values. The synonymous substitution rate (*Ks*) was also calculated using MYN algorithm(Wang *et al.*, 2009) based on the Tamura-Nei Model to further confirm 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 Lab (Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences) was characterized on a Bruker AV 400 spectrometer using tetramethylsilane (TMS, $\delta = 0$) as internal reference. The spectra was consistent with the data in the reference (Figure S3, Yuan *et al.*, 2020).

All the intermediates were detected by UPLC-Q-TOF （UPLC1290-6540B Q-TOF, Agilent, USA) . The mobile phase comprised with 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, Japan). 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 sample tray 4°C. The standard substances included secologanin (Absin, China, 19351-63-4), tryptamine (Sigma, USA, 61-54-1), 3 α -dihydrocadambine and epoxystrictosidine (Nature Standard, China, 54483-84-0).

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

The coding sequence of *evm.model.contig69.90* without start codon was amplified with a primer pairs (forward primer: 5'-CTCTTGGAAATCATGCCTCACA- 3' and reverse primer: 5'- TCAGACAGAAGAAACCACTCCATTC-3') and cloned into the pEASY®- Blunt E1 Expression Vector (TransGen Biotech,Beijing, China), 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 ℃ for a given time (0 hr, 24 hr, 48 hr). The reaction was terminated by 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 genes 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's (Qiagen, USA) protocol. Clean RNA-seq reads were mapped to *N. cadamba* reference genome using Hisat2, then the expression level for *N. cadamba* genes (FPKM, and expression count data) was obtained using HTSeq software(Kim *et al.*, 2015b, Kim *et al.*, 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, USA) as described in the previous method (LC-MS/MS). GWAS analyses were performed in the mixed linear model using GEMMA (v 0.94.1)(Zhou and Stephens, 2012). The Bonferroni correction was used to reduce the probability of false positives. And, $P = 10^{-5}$ were was used as the genome-wide thresholds to screen significantly associated sites. The candidate genes were obtained from 5000bp region upstream and downstream of a peak value of significant region.

Reads mapping and variant calling.

The clean reads were mapped back to the assembled genome using BWA-MEM (v 0.7.8-r455) (Li and Durbin, 2009) with modified parameters: the minimum seed length was set to 32, band width for banded alignment was set to 150 and penalty for a mismatch was set to 3. PCR duplicates were removed using the rmdup function of SAMtools (v 0.1.19-44428cd)(Li *et al.*, 2009) with default parameters. SNPs were detected using SAMtools (v 0.1.19-44428cd) (Li *et al.*, 2009) and bcftools (v 1.3.1)(Narasimhan *et al.*, 2016). The obtained SNPs were further filtered when the

mapped reads <4, the missing data > 0.3 or the Minor Allele frequencies < 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, USA) as described in the previous method (LC-MS/MS). All the filtered SNPs were used to construct a neighbor-joining phylogenetic tree using TreeBeST (v 1.9.2) with 1000 bootstraps and other parameters default. The phylogenetic tree was visualized using Evolview v2 (https://www.evolgenius.info//evolview/). In addition, the principal component analysis (PCA) was performed using the same SNPs set by GCTA (v 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 a 20 kb steps for each population by VCFtools (v 0.1.14). The selective regions were detected by the top 5% of the low or high $\log_2 \pi$ ratio and the top 5% of *F*st values. The linkage disequilibrium (LD) of each population was estimated by calculating squared correlation coefficient (r^2) values between any two SNP sites in 500 kb using Haploview (v 4.2)(Barrett *et al.*, 2005).

Data availability

All raw and processed sequencing data generated in this study have been submitted to the NCBI BioProject database under accession number PRJNA650253. The raw genome sequencing data obtained by Illumina and PacBio platform 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 number SAMN15700860 and SAMN15700859, respectively. The genome annotation and assembled genome sequences can be downloaded from the webpage (https://figshare.com/s/ed20e0e82a4e7474396b).

Author contributions

C.P., X.Z. and X.C. designed and managed the project. C.P., X.Z. and X.H. wrote the manuscript with input from all authors. K.O., Q.Q., J. L. and J. Z. collected samples. T.Z. and J.G. performed the *in vitro* catalytic assay. J.Y., T.Z. and X.H. performed HPLC-MS/MS experiments. K.O., X.W., S.Y. and L.Z. performed RNA-seq analyses of gene expression. S.L. and W.G. performed FISH analysis; X.H., W.J., B. L., K.O. and X. Z. worked on sequencing and data analyzing. C.P., X.Z., X.H., E.N., X.C. and K.O. revised the manuscript. All the authors read and approved the final manuscript.

Acknowledgements

We thank Y. Shang for critical reading of the manuscript. 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 No.31470681, 31970197); the National Key Research and Development Program of China (Grant No.2016YFD0600104); Natural Science Foundation of Guangdong Province of China (Grant No. 2016A030311032); Guangzhou Science and Technology Program (Grant No. 201607020024) and Foundation of Young Creative Talents in Higher Education of Guangdong Province (Grant No.2017KQNCX017).

Conflict of interest

The authors declare no competing interests.

Supplemental Figures

Figure S1. A. 17 K-mer analysis for estimating the genome size of *Neolamarckia cadamba* B. Cytological analysis of *Neolamarckia cadamba* metaphase chromosomes by FISH using 45s rDNA as 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 (*Ks*) in *N. cadamba* genome

Figure S3 Characterization of standard compounds by Q-TOF LC-MS/MS and NMR. Tryptamine, epoxystrictosidine, 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 CD_3OD 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 epoxystrictosidine (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 S 5. *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 predicted epoxystrictosidine.

Figure S7 The expression profile of all the predicted biosynthetic genes in Shikimate pathway.

Figure S8 The expression profile of all the predicted squalene epoxidase genes (SQEs).

Figure S9. Phylogenetic tree of NcCYPs 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* paralogues.

Figure S11 The structure of tandem duplicated NcDLGTs.

Figure S12 Heat Map of Expression Data of genes in NcMCL24 and NcMCL25.

Figure S13 Duplication of NcMYC TFs in tandem order.

Figure S14 Protein alignment of CrSGD, RsSGD, RvSGD, RsRG and their respective *Neolamarckia* and Ophiorrhiza paralogues. 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.

Supplemental Tables

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

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

Table S3. Coverage statistics of *Neolamarckia cadamba* genome

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

Table S5. Assessment the genome assembly using RNA-seq

Table S6. The statistical results of gene function annotation of *Neolamarckia cadamba* genome.

Table S7. The statistical results of non-coding RNA of *Neolamarckia cadamba* genome.

Table S8. Summary of Repeat contents in *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 Gene lists under positive selection.

Table S12 The predicted biosynthetic genes in five plant species.

Table S13. Short lists of *Neolamarckia cadamba* candidate genes involved in MVA/MEP,

shikimate/indole pathway and iridoid biosynthesis pathway and 35 putative NcSTRs.

Table S14 Unique gene lists underwent recent WGD event.

Table S 15 Gene lists in the association network of gene-to-metabolite.

Table S 16 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 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 plantiSMASH algorithm

Table S21 The 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): 1469–84.

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. *Plant Cell,* 19**,** 2886-97.

BARRETT, J. C., FRY, B., MALLER, J. & DALY, M. J. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics,* 21**,** 263-5.

BEDI, Y. S., BIR, S. S. & GILL, B. S. 1981. Cytopalynology of Woody Taxa of Family Rubiaceae from North and Central India. P*roceedings of the indian national science academy.part b.biological sciences*.47 (6):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,* 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., DUGE DE BERNONVILLE, T., OUDIN, A., LANOUE, A., BILLET, K., MUNSCH, T., KOUDOUNAS, K., MELIN, C., GODON, C., RAZAFIMANDIMBY, B., DE CRAENE, J. O., GLEVAREC, G., MARC, J., GIGLIOLI-GUIVARC'H, N., CLASTRE, M., ST-PIERRE, B., PAPON, N., ANDRADE, R. B., O'CONNOR, S. E. & COURDAVAULT, V. 2018a. Two Tabersonine 6,7-Epoxidases Initiate Lochnericine-Derived Alkaloid Biosynthesis in Catharanthus roseus. *Plant Physiol,* 177**,** 1473-1486.

CARQUEIJEIRO, I., DUGE DE BERNONVILLE, T., LANOUE, A., DANG, T. T., TEIJARO, C. N., PAETZ, C., BILLET, K., MOSQUERA, A., OUDIN, A., BESSEAU, S., PAPON, N., GLEVAREC, G., ATEHORTUA,

L., CLASTRE, M., GIGLIOLI-GUIVARC'H, N., SCHNEIDER, B., ST-PIERRE, B., ANDRADE, R. B., O'CONNOR, S. E. & COURDAVAULT, V. 2018b. A BAHD acyltransferase catalyzing 19-O-acetylation of tabersonine derivatives in roots of Catharanthus roseus enables combinatorial synthesis of monoterpene indole alkaloids. *Plant J,* 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. *Perspectives in Cancer Prevention-Translational Cancer Research.* Springer.

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., DUNN, C., O'MALLEY, R., FIGUEROA-BALDERAS, R. & MORALES-CRUZ, A. 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 Plant Sci,* 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**,** 1269-71.

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. & LU, L. 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., ZHENG, C. F., ALBERTI, A., ANTHONY, F., APREA, G., AURY, J. M., BENTO, P., BERNARD, M., BOCS, S., CAMPA, C., CENCI, A., COMBES, M. C., CROUZILLAT, D., DA SILVA, C., DADDIEGO, L., DE BELLIS, F., DUSSERT, S., GARSMEUR, O., GAYRAUD, T., GUIGNON, V., JAHN, K., JAMILLOUX, V., JOET, T., LABADIE, K., LAN, T. Y., LECLERCQ, J., LEPELLEY, M., LEROY, T., LI, L. T., LIBRADO, P., LOPEZ, L., MUNOZ, A., NOEL, B., PALLAVICINI, A., PERROTTA, G., PONCET, V., POT, D., PRIYONO, RIGOREAU, M., ROUARD, M., ROZAS, J., TRANCHANT-DUBREUIL, C., VANBUREN, R., ZHANG, Q., ANDRADE, A. C., ARGOUT, X., BERTRAND, B., DE KOCHKO, A., GRAZIOSI, G., HENRY, R. J., JAYARAMA, MING, R., NAGAI, C., ROUNSLEY, S., SANKOFF, D., GIULIANO, G., ALBERT, V. A., WINCKER, P. & LASHERMES, P. 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 Res,* 32**,** 1792-7.

EGER, E., SIMON, A., SHARMA, M., YANG, S., BREUKELAAR, W. B., GROGAN, G., HOUK, K. & KROUTIL, W. 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., WU, H., QIN, X., YAN, L., TAN, L., SIM, S., LI, W., SASKI, C. A., DANIELL, H., WENDEL, J. F., LINDSEY, K., ZHANG, X., HAO, C. & JIN, S. 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., CEPELA, J., HABERMANN, M., STEUERNAGEL, B. & CLISSOLD, L. 2015. Genome‐guided investigation of plant natural product biosynthesis. *The Plant Journal,* 82**,** 680-692.

KETUDAT CAIRNS, J. R. & ESEN, A. 2010. β-Glucosidases. *Cell Mol Life Sci,* 67**,** 3389-405.

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. *Nat Methods,* 12**,** 357-60.

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 Biol,* 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**,** 1754-60.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & SUBGROUP, G. P. D. P. 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***,** 18.1119 a.

LI, L., STOECKERT, C. J., JR. & ROOS, D. S. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res,* 13**,** 2178-89.

Li, M., Tian, S., Jin, L., Zhou, G., Li, Y., Zhang, Y., … Ma, J. 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 Plant Sci,* 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. *J Biol Chem,* 280**,** 13576-83.

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.P1ant 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**,** 1749-51.

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**,** 1061-7.

PATERSON, A. H., BOWERS, J. E. & CHAPMAN, B. A. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci U S A,* 101**,** 9903-8.

PRESSNITZ, D., FISCHEREDER, E. M., PLETZ, J., KOFLER, C., HAMMERER, L., HIEBLER, K., LECHNER, H., RICHTER, N., EGER, E. & KROUTIL, W. 2018. Asymmetric Synthesis of (R) ‐ 1 ‐ Alkyl ‐ Substituted Tetrahydro ‐ ß ‐ carbolines Catalyzed by Strictosidine Synthases. *Angewandte Chemie,* 130**,** 10843-10847.

PUTTICK, M. 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. & DE LUCA, V. 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,* 115**,** 3180-3185.

RAI, A., HIRAKAWA, H., NAKABAYASHI, R., KIKUCHI, S., HAYASHI, K., RAI, M., TSUGAWA, H., NAKAYA, T., MORI, T. & NAGASAKI, H. 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 Data. A New Classi. *Syst.geogr.pl,* 76**,** 85-145.

ROSSELEENA, R. E., MOTOAKI, C., MIKI, K., TAKASHI, A., YOSHIMI, O., NOBUTAKA, M., MASARU, O. T., ATSUSHI, F., AMIT, R. & KAZUKI, S. 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., PARK, S., SABIR, M. J., BAESHEN, M. N. & HAJRAH, N. H. 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. & DELLAPENNA, D. 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. G., K.O.G., N., COSTA E.V., SOUZA A.D.L., M.L.B., P., H.H.F., K. & F.M.A, S. 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. 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**,** 3210-2.

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, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics,* 30**,** 1312-3.

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., CA RQUEIJEIRO, I., THOMAS, D., FRANKE, J., DANG, T. T. T., OUDIN, A., LANOUE, A., LAFONTAINE, F., STAVRINIDES, A. K. & CLASTRE, M. 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**,** 1105-11.

TRAPNELL, C., ROBERTS, A., GOFF, L., PERTEA, G., KIM, D., KELLEY, D. R., PIMENTEL, H., SALZBERG, S. L., RINN, J. L. & PACHTER, L. 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., MIETTINEN, K., VANDEN BOSSCHE, R., DE CLERCQ, R., MEMELINK, J. & GOOSSENS, A. 2016. The basic helix-loop-helix transcription factor BIS2 is essential for monoterpenoid indole alkaloid production in the medicinal plant Catharanthus roseus. *Plant J,* 88**,** 3-12.

VAN MOERKERCKE, A., STEENSMA, P., SCHWEIZER, F., POLLIER, J., GARIBOLDI, I., PAYNE, R., VANDEN BOSSCHE, R., MIETTINEN, K., ESPOZ, J., PURNAMA, P. C., KELLNER, F., SEPPANEN-LAAKSO, T., O'CONNOR, S. E., RISCHER, H., MEMELINK, J. & GOOSSENS, A. 2015. The bHLH transcription factor BIS1 controls the iridoid branch of the monoterpenoid indole alkaloid pathway in Catharanthus roseus. *Proc Natl Acad Sci U S A,* 112**,** 8130-5.

WANG, C., WU, C., WANG, Y., XIE, C., SHI, M., NILE, S., ZHOU, Z. & KAI, G. 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 (1)： 20-20

WANG, Y., TANG, H., DEBARRY, J. D., TAN, X., LI, J., WANG, X., LEE, T. H., JIN, H., MARLER, B., GUO, H., KISSINGER, J. C. & PATERSON, A. H. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res,* 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., GAO, X., SONG, L. D., LI, Y. & PENG, L. Y. 2013. Two new indole alkaloids from *Emmenopterys henryi*. *Helvetica Chimica Acta,* 96**,** 2207-2213.

XIA, L., RUPPERT, M., WANG, M., PANJIKAR, S., LIN, H., RAJENDRAN, C., BARLEBEN, L. & STOCKIGT, J. 2012. Structures of alkaloid biosynthetic glucosidases decode substrate specificity. *ACS Chem Biol,* 7**,** 226-34.

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 Cell Physiol. 44, 395–403

YANG, J., LEE, S. H., GODDARD, M. E. & VISSCHER, P. M. 2011. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet,* 88**,** 76-82.

YANG, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol,* 24**,** 1586-91.

YUAN, H.-L., ZHAO, Y.-L., QIN, X.-J., LIU, Y.-P., YU, H.-F., ZHU, P.-F., JIN, Q., YANG, X.-W. & LUO, X.-D. 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., GANTET, P. & MEMELINK, J. 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. *Plant J,* 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. *Nat Methods,* 11**,** 407-9.

Table 1. Global statistics of *N. 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)		3,489.6 bp
Average CDS length (bp)		1,151.7 bp
Average exons per gene	4.7	
Average exon length (bp)		245.1 bp
Average intron length (bp)		632.0 bp

Figure 1 Hi-C map and overview of the genomic features of the *N. cadamba* 22 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: (i) Circular representation of the 22 chromosomes. (ii-vi) Densities of transposable element, gene, GC, miRNA and tRNA. (vii) Densities of snRNA density and syntenic blocks. Densities were calculated in 100Kb 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), *C. canephora* (Cca) and *S. tuberosum* (Stu). **e** Synteny blocks between *N. cadamba*, *C. canephora* and *O. pumila*.

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 methods)**. a** and **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 in *N. cadamba* by LC-MS. **d** and **e** Expoxystrictosidine, 3α-dihydrocadambine and cadambine were detected in the barks of *N. cadamba* by LC-MS. **f- j** MS/MS chromatograms indicated those substances represented by the peak in the previous Fig. c-e as tryptamine (**f**, 3.62 min), strictosidine (**g**, 13.02 min), 3α-dihydrocadambine (**h**,

12.30 min), expoxystrictosidine (**i**, 13.57min) and cadambine (**j**, 12.33 min), respectively.

Figure 3 Neolamarckia strictosidine synthase gene candidates selection and functional identification of NcSTR1. **a** Alignment of the active domains in STRs from *Ophiorrhiza pumila* (Op), *Catharanthus roseus* (Cr), *Rauvolfia serpentine* (Rs) and the five Neolamarckia homologous with the highest similarity to CrSTR at amino acid level. Previously reported active-site residues of OpSTR (Q94LW9 ， Eger,2020）,CrSTR (CAA43936.1 ， Petronikolou, 2018) and RsSTR (P68175.1， Ma 2006) were indicated in red color letters under the aligned sequences. The essential active site glutamate residue (Glu309 in RsSTR1) were highly conserved in NcSTR1, NcSTR13 and NcSTR14. **b** The expression profile of all 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 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 curve in blue and pink indicated the 531.10 ->352.20 and 531.10 ->514.25 ion pairs respectively, and black indicated the superimposed curve with blue and pink curve. (i) The standard of strictosidine. (ii)The negative control without the enzyme NcSTR1 48 hours after the reaction. (iii) The NcSTR1 catalyzed 24 hours after the reaction. (iv) The NcSTR1 catalyzed 48 hours after the reaction. (v)The catalyzed Pictet-Spengler chemical reaction by NcSTR1.

Figure 4 The predicted cadambine biosynthetic pathway and MIA biosynthetic gene candidates. a Proposed cadambine biosynthetic pathway in plants. Dotted line arrows indicate multiple steps between intermediates. **b** The expression profile of all the 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

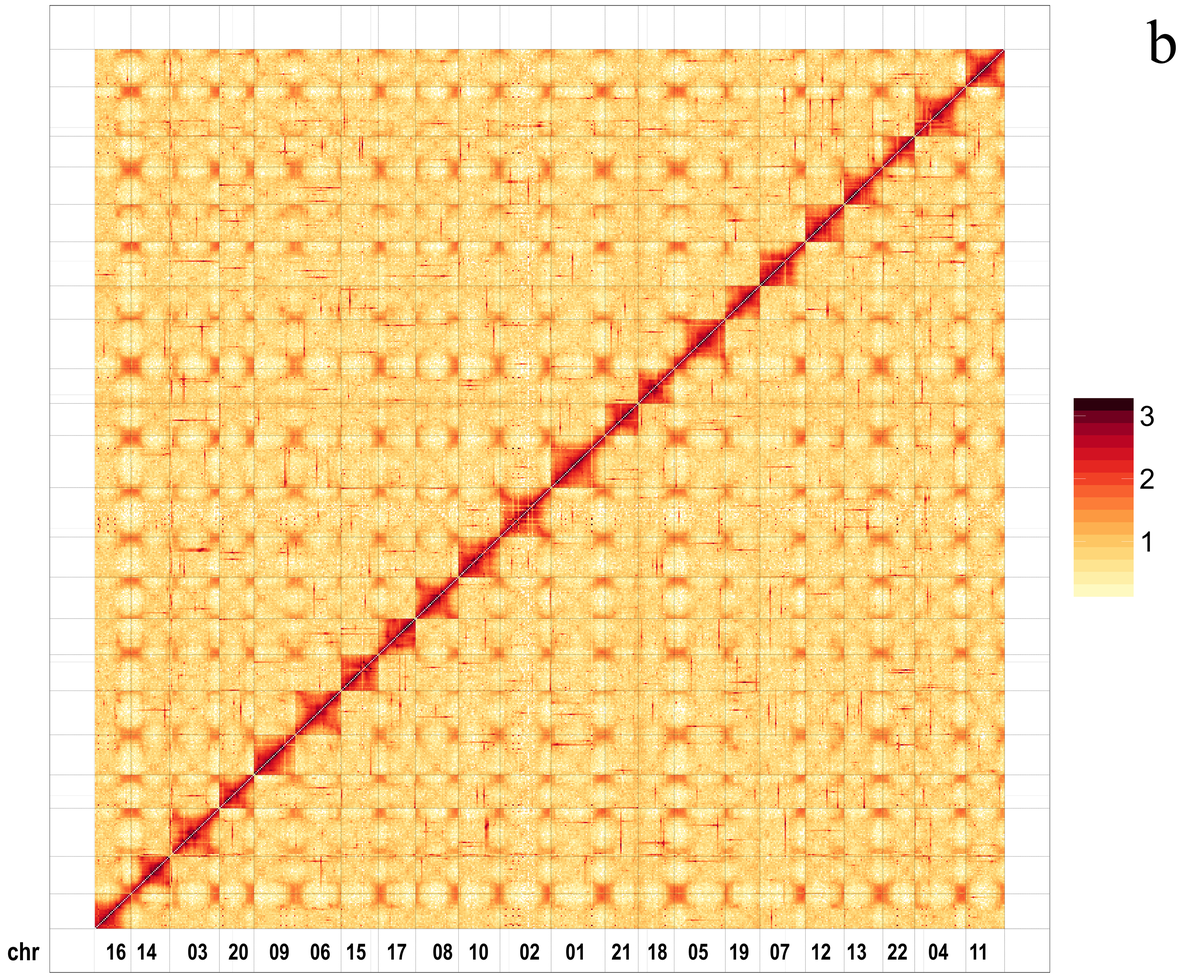
(transitional cambium, TCA; secondary, SCA) and phloem (primary phloem, PPH; transitional phloem, TPH; secondary phloem, SPH) from the first, second and fourth internodes. Enzymes in abbreviations are: 7-DLH, 7-deoxyloganic acid hydroxylase (CYP72A224); 8-HGO, 8-hydroxy-geraniol oxidoreductase; AS, anthranilic 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 beta-glucosidase; SLS, Secologanin synthetase; STR, strictosidine synthase; TDC, tryptophan decarboxylase; THAS, tetrahydroalstonine synthase (Kai et al. 2015; Wu et al.2018; Rai, 2021; Qu, 2019; Yang,2019).

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 LD, measured by r2, 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 were highlighted with red arrow.

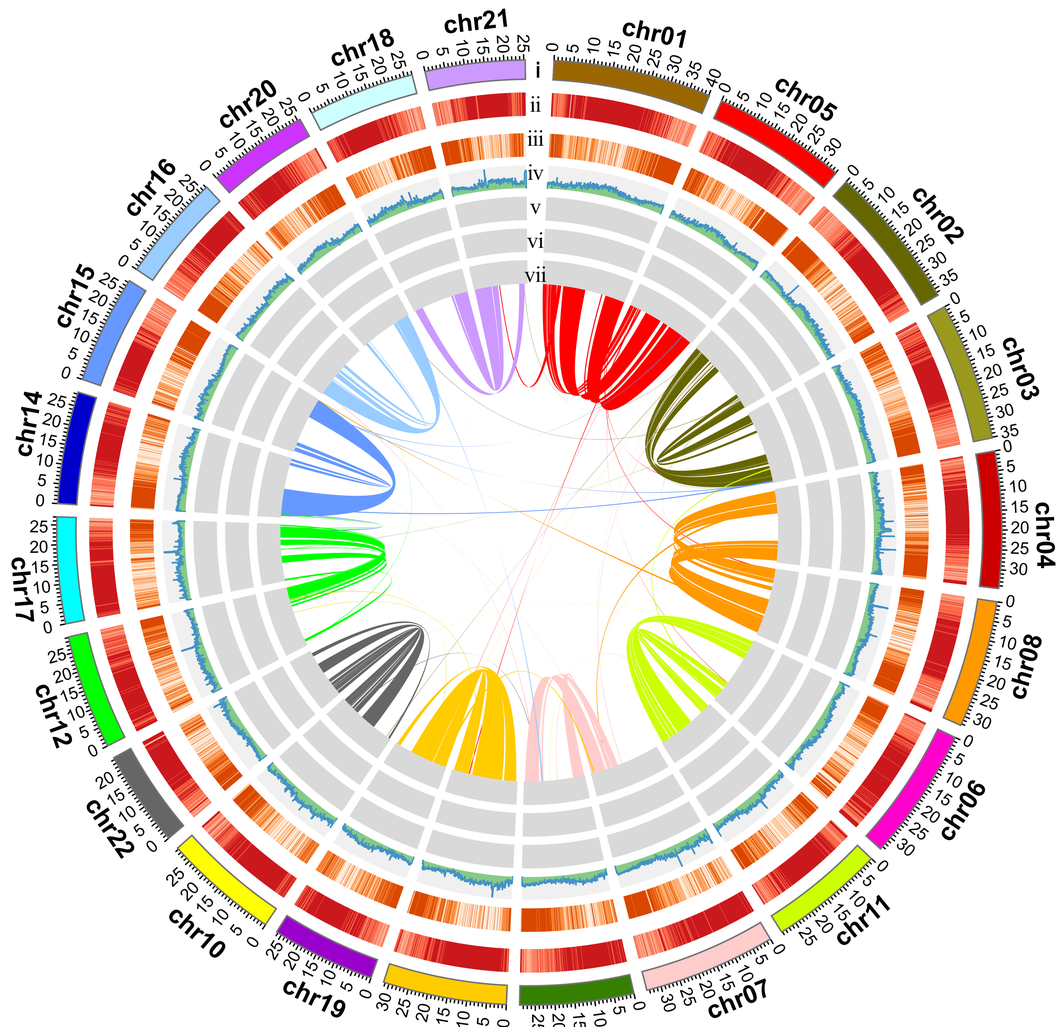
Figure 6 The structure and evolution of the predicted cadambine biosynthetic gene clusters and tandem duplications. a Schematic diagram of the predicted gene clusters and tandem duplicates in *N. cadamba.* NcSMASH35: TDC/STR1/CYPs; NcMCL22: CYP716/NcSTR14/BGLUs; NcMCL18: GES/NcSTR13/CYPs; NcSMASH9, NcSMASH65 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 bHLH TFs in NcMCL25 and NcMCL24. Enzymes in abbreviations are: 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 paralogue; 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.

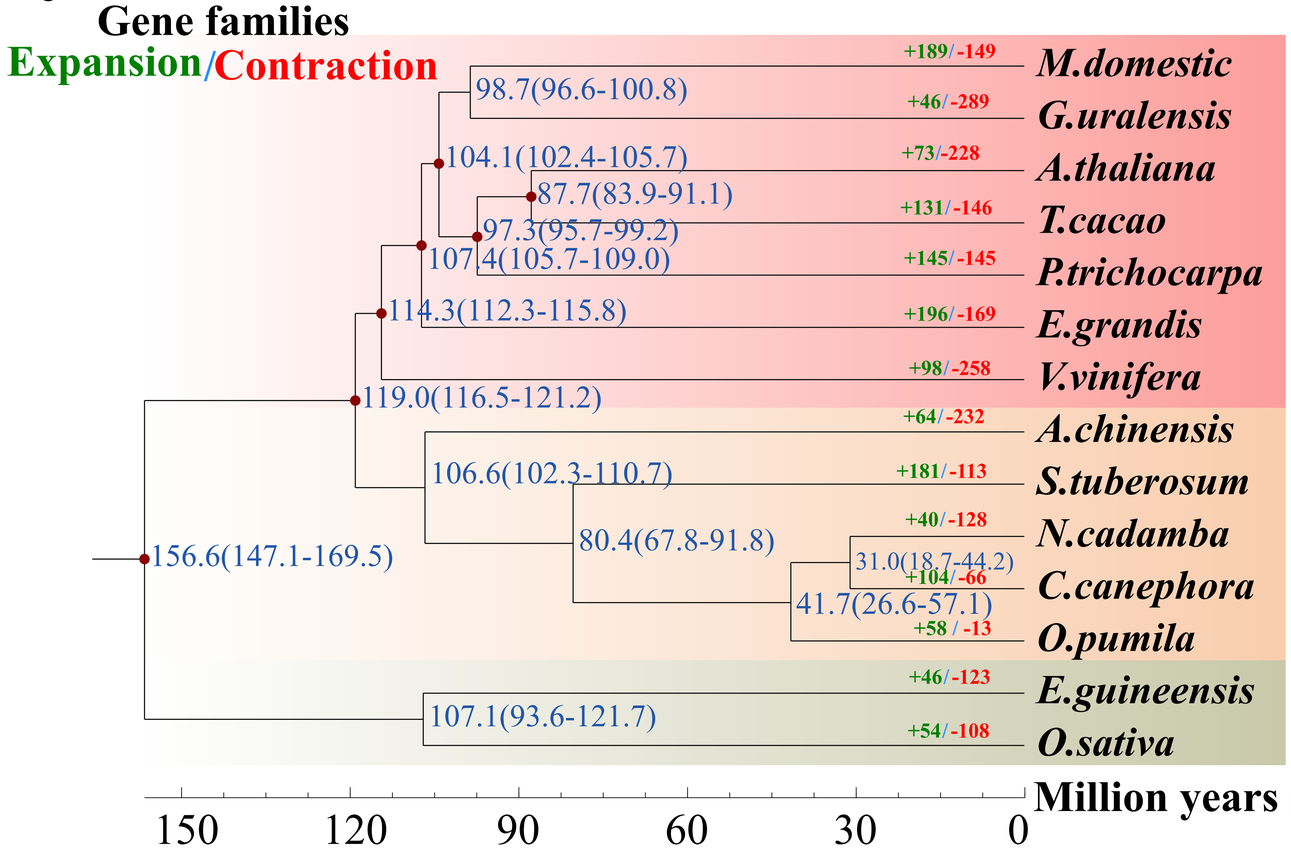
a



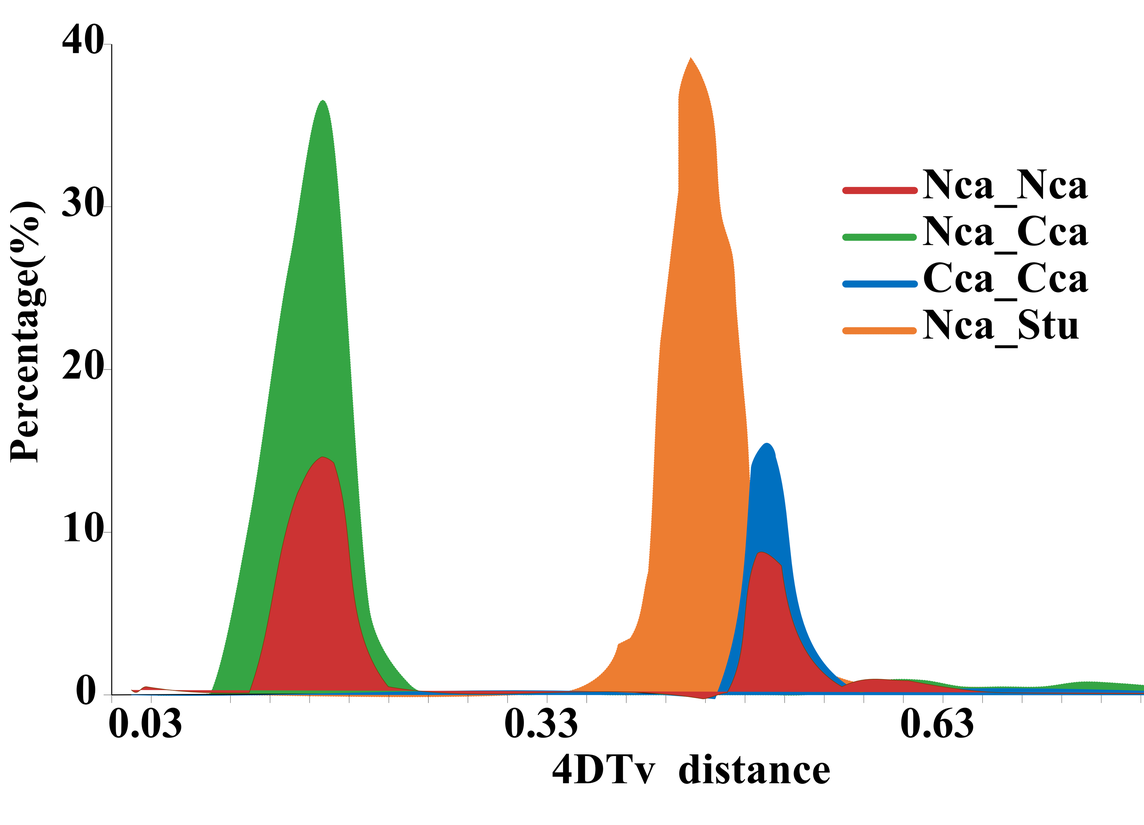
b



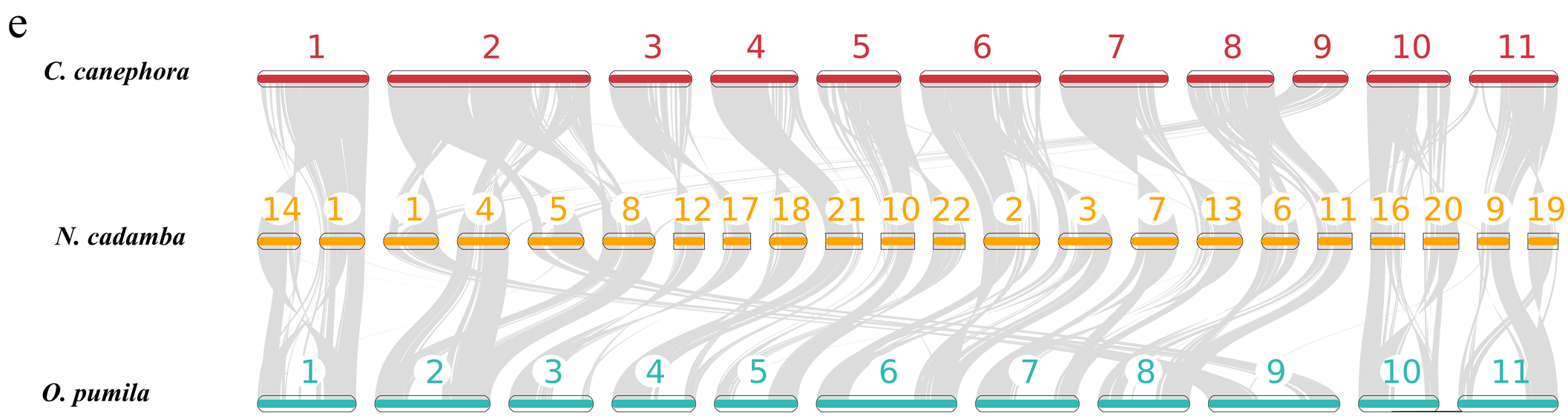
c

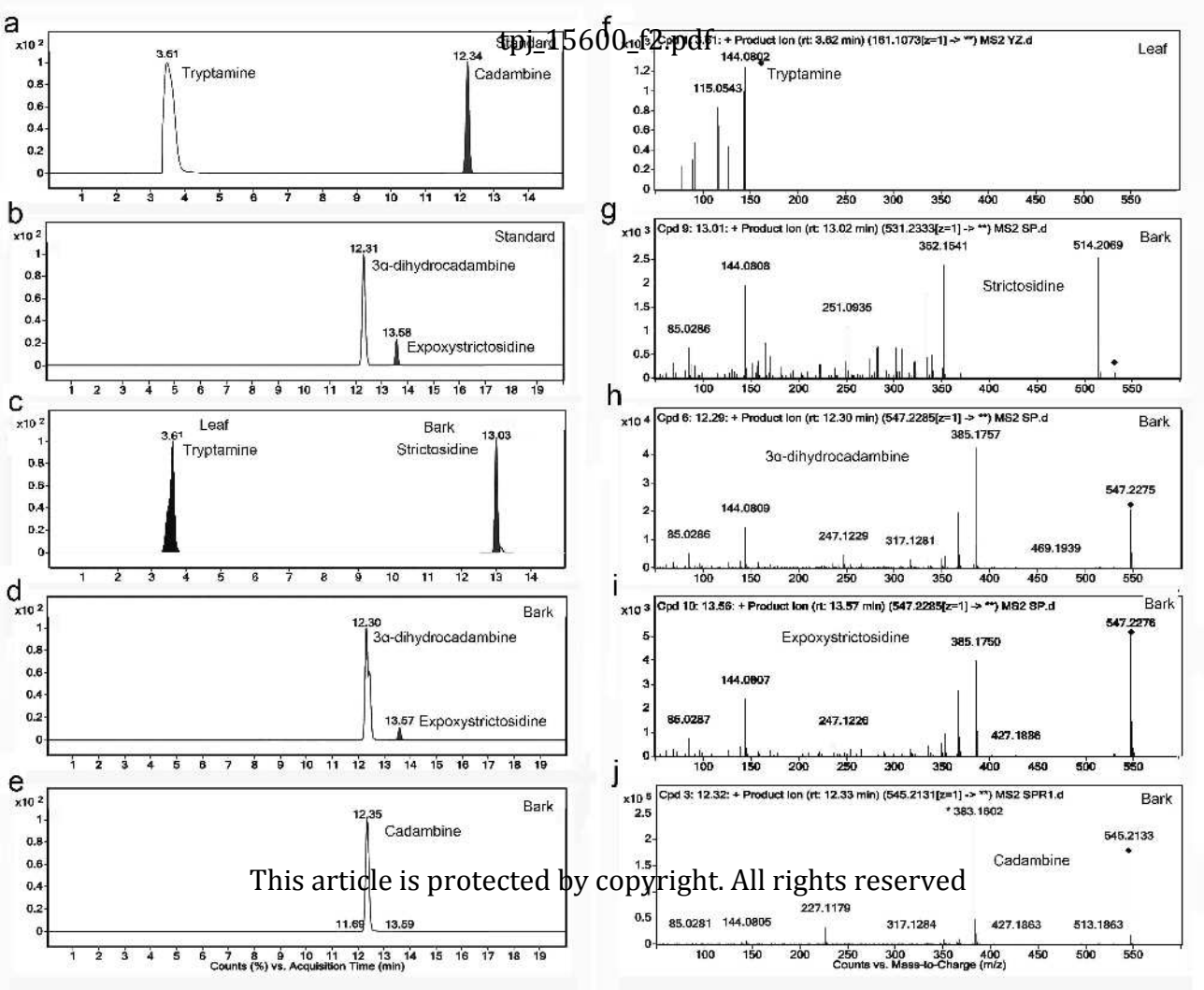


d



e

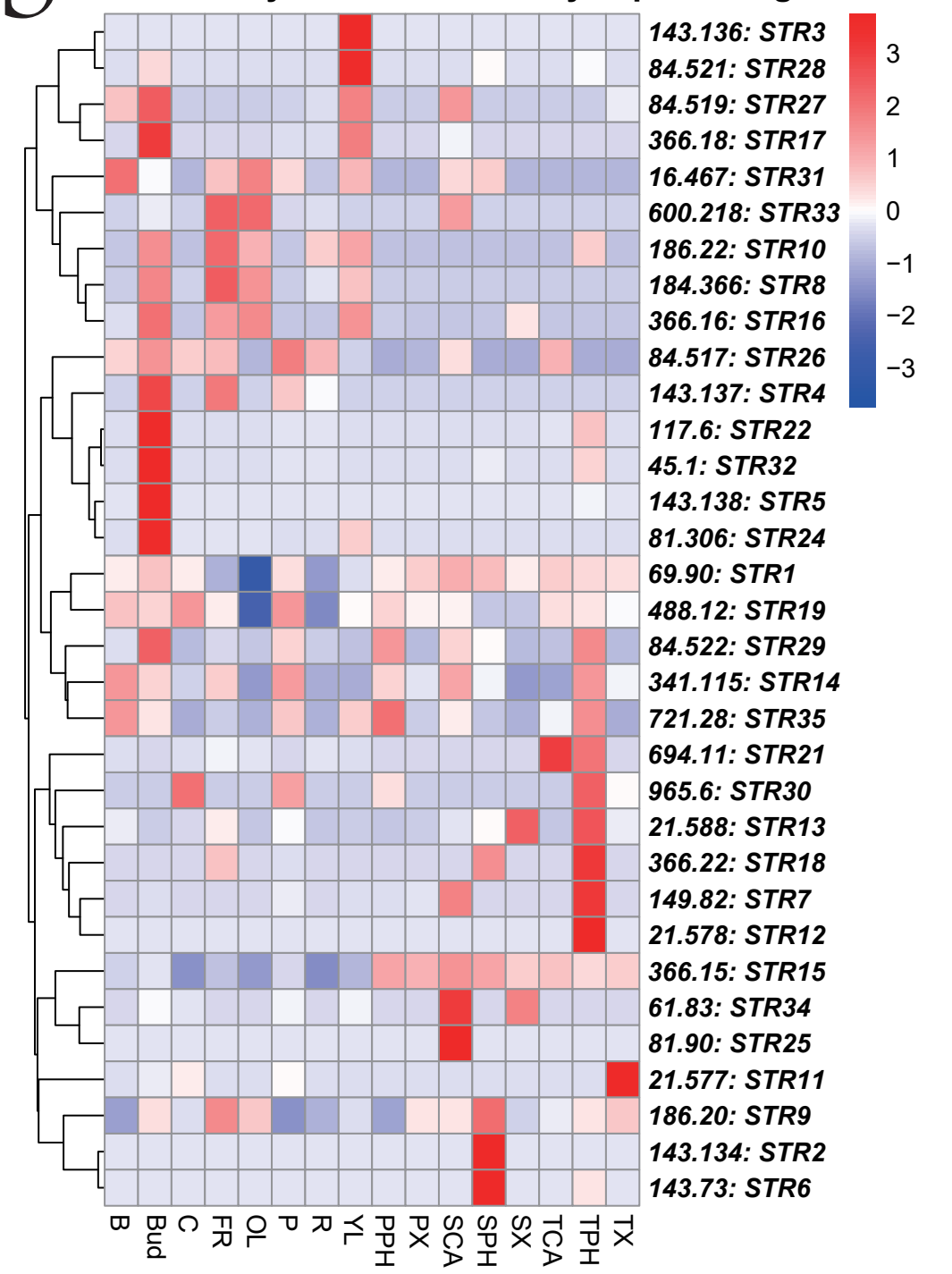




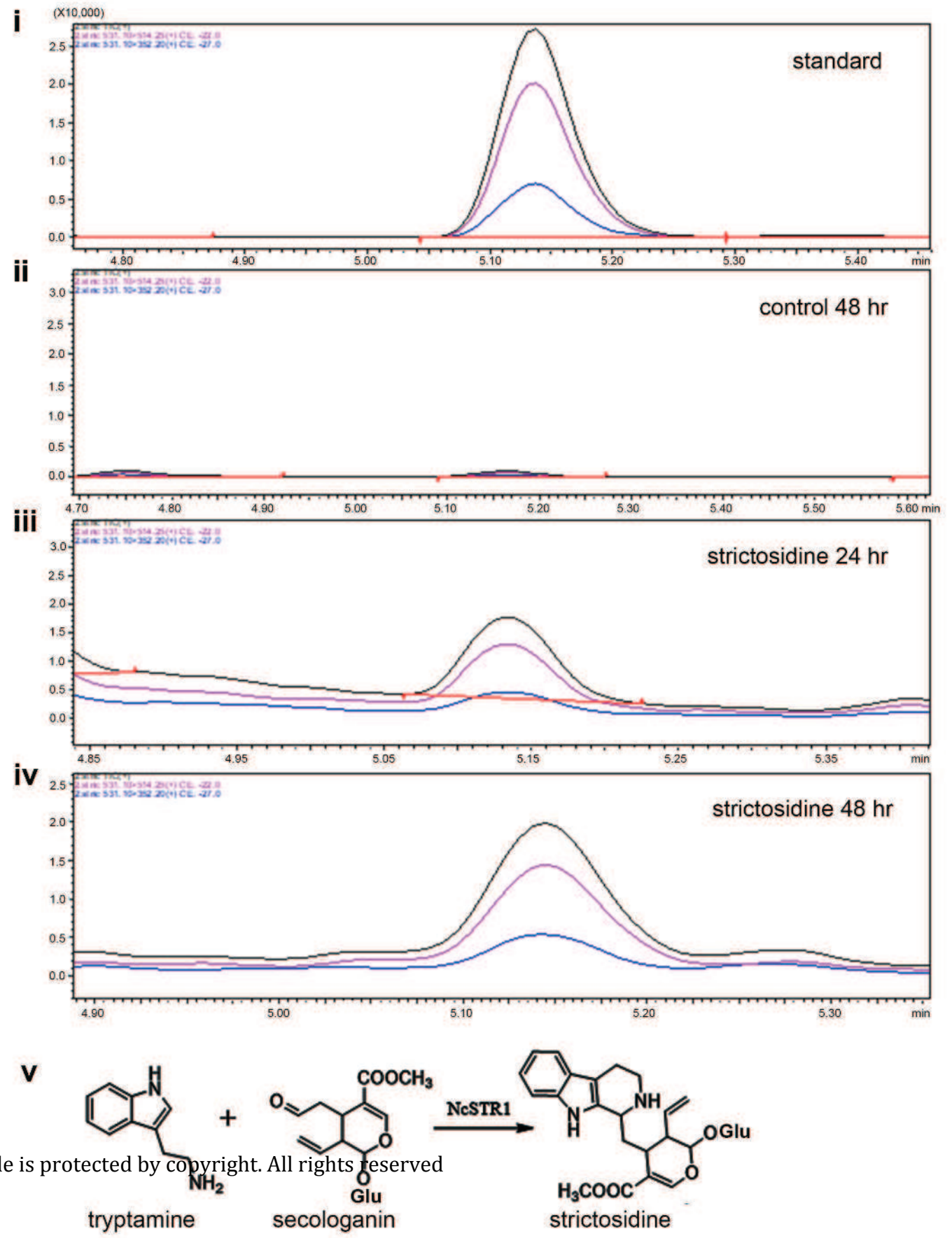
a

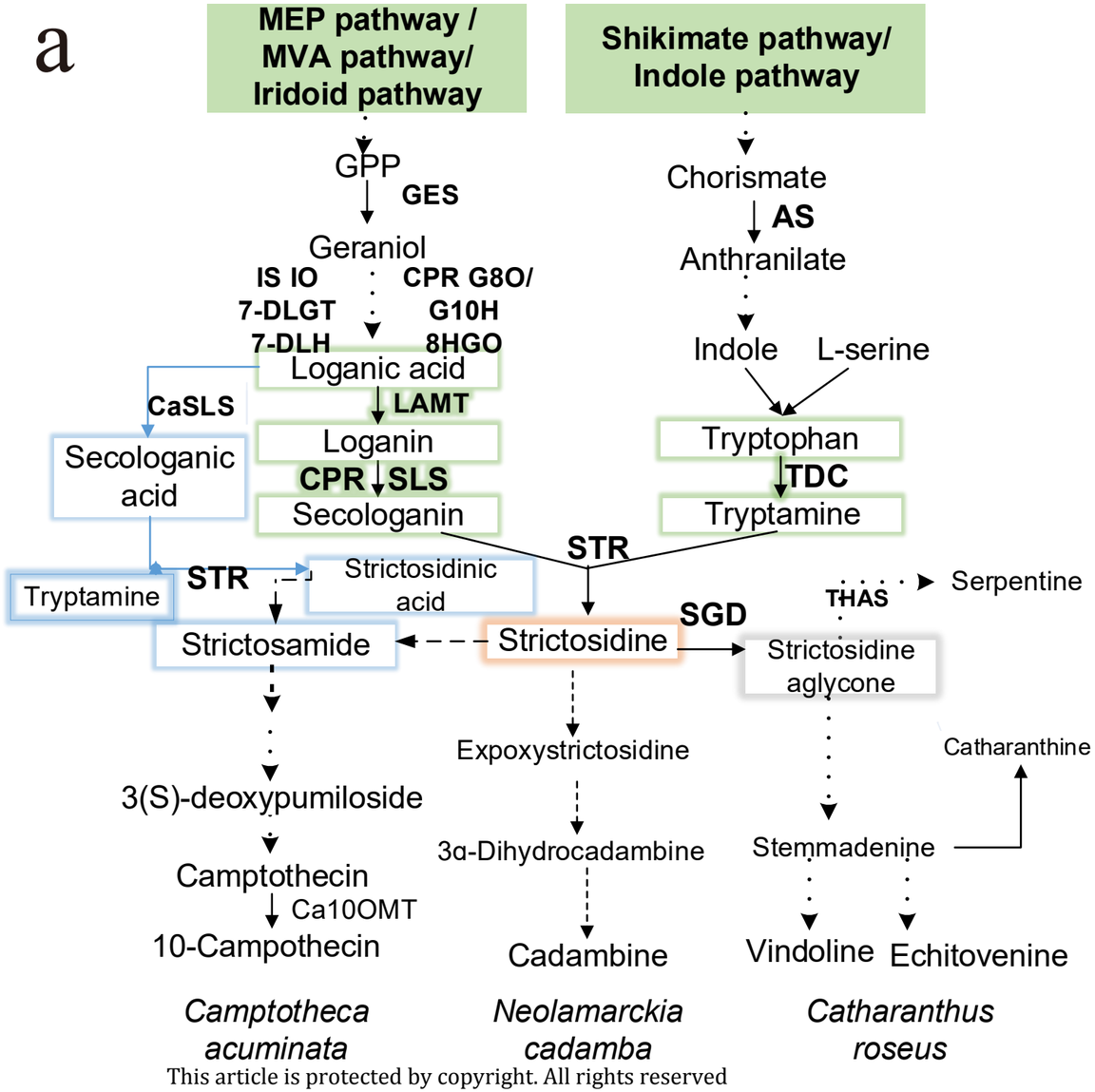
```
                    80          *         100
NcSTR22  :  VKYAGSDIGVVNFSYASPFRTAKFCDGTTDPQKGY  :  85
NcSTR13  :  VKYKGSSNGFSTHAVASPIWNKEVCENNTNPQLKP  :  89
NcSTR14  :  VKYKGSSNGFSTHAVASPIWNKEVCENNTNPQLKP  :  89
NcSTR 1  :  VKYEGSRNGFLTHAVPSPVWSTKVCENNSKPELKP  : 101
NcSTR27  :  LKYESPEIGFVEFAYASPIRTREECDGTTNPDLGP  :  90
Op__STR  :  IKYDKPSNKFLTHAVASPIWNNALCENNTNQDLKP  :  95
Cr__STR  :  IKYEGPNSGFTDFAYASPFWNKAFCENSTDPEKRP  : 103
Rs__STR  :  KY PNS FVDF Y SP NKAFCENS DA K P  :  99
            KY G   GF   A ASP W   C89N91 T   P
                        *       160       *
NcSTR22  :  ATLLSNSAGGRPYKWLDGVDVDQTTGKIYFTDISA  : 155
NcSTR13  :  ATQLAKSADGVDFKWLYALAVDQQTGFVYLTDVST  : 159
NcSTR14  :  ATQLAKNADGVDFKWLYALAVDQQTGFVYLTDVST  : 159
NcSTR 1  :  ATQVAKGADGRDFKWLYALAVDQETGSVYTTDVSA  : 171
NcSTR27  :  AKQLATSAEGCPFKWLDALDVDSRTGIVYFTDIST  : 160
Op__STR  :  AIQLATSGDGVEFKWLYALAIDQQAGFVYVTDVST  : 165
Cr__STR  :  ATQLATSVQGVPFKWLYAVTVDQRTGIVYFTDVSS  : 173
Rs__STR  :  ATQLATSVDGVPFKWLYAVTVDQRTGIVYFTDVST  : 169
        ATQLA S  G   FKW145LY147AL VDQ TG VY TD S
                    220         *         240
NcSTR22  :  VSTGLQGPGGVGASLDGSFVLYSEFVGNKITKYWL  : 223
NcSTR13  :  MN-GLNVPGGTEVSKDGSFVVVAEFLSHRILKYWL  : 226
NcSTR14  :  MN-GLNVPGGTEVSKDGSFVVVAEFLSHKILKFWL  : 226
NcSTR 1  :  LK-DLNIPGGVELGEDGSFLLVGEYLSHRILKYWL  : 238
NcSTR27  :  MS-GIPGPGGIAVSKDGSFILVSEYFGDRILKYWL  : 227
Op__STR  :  MK-GLNIPGGTEVSKDGSFVLVGEFASHRILKYWL  : 232
Cr__STR  :  LK-ELHVPGGAEISADGSFVVVAEFLSNRIVKYWL  : 242
Rs__STR  :  LK-ELHVPGGAEVSADSSFVLVAEFLSHQIVKYWL  : 236
               L  PGG E S  DSFV V  EF222 S  I KYWL
            320         *         340         *
NcSTR22  :  GTFITEVHEHRGKLYIASMYQ-LPGVI--YT-LPD  : 320
NcSTR13  :  GEHFEQVQEHNGALYIGSLFHDFVGILHNYEGSSD  : 325
NcSTR14  :  GEHFEQVQEHNGALYIGSLFHDFVGILHNYEGSSN  : 325
NcSTR 1  :  GEFYEQVQEYNGALYIGSLFHDFVGILHNYDGSSD  : 337
NcSTR27  :  NSVINEVQEFGDSYYIGSLY------LP-FAAASK  : 322
Op__STR  :  GEHIEQVQEHDGALFVGSLFHEFVGILHNYKSSVD  : 331
Cr__STR  :  GEHFEQIQEHDGLLYIGSLFHSSVGILV-YD-DHD  : 343
Rs__STR  :  GEHFEQIQEHDGLLYIGTLFHGSVGILV-Y----D  : 334
        GEH307E309QQE  G L  G  FH  VG  L Y
```

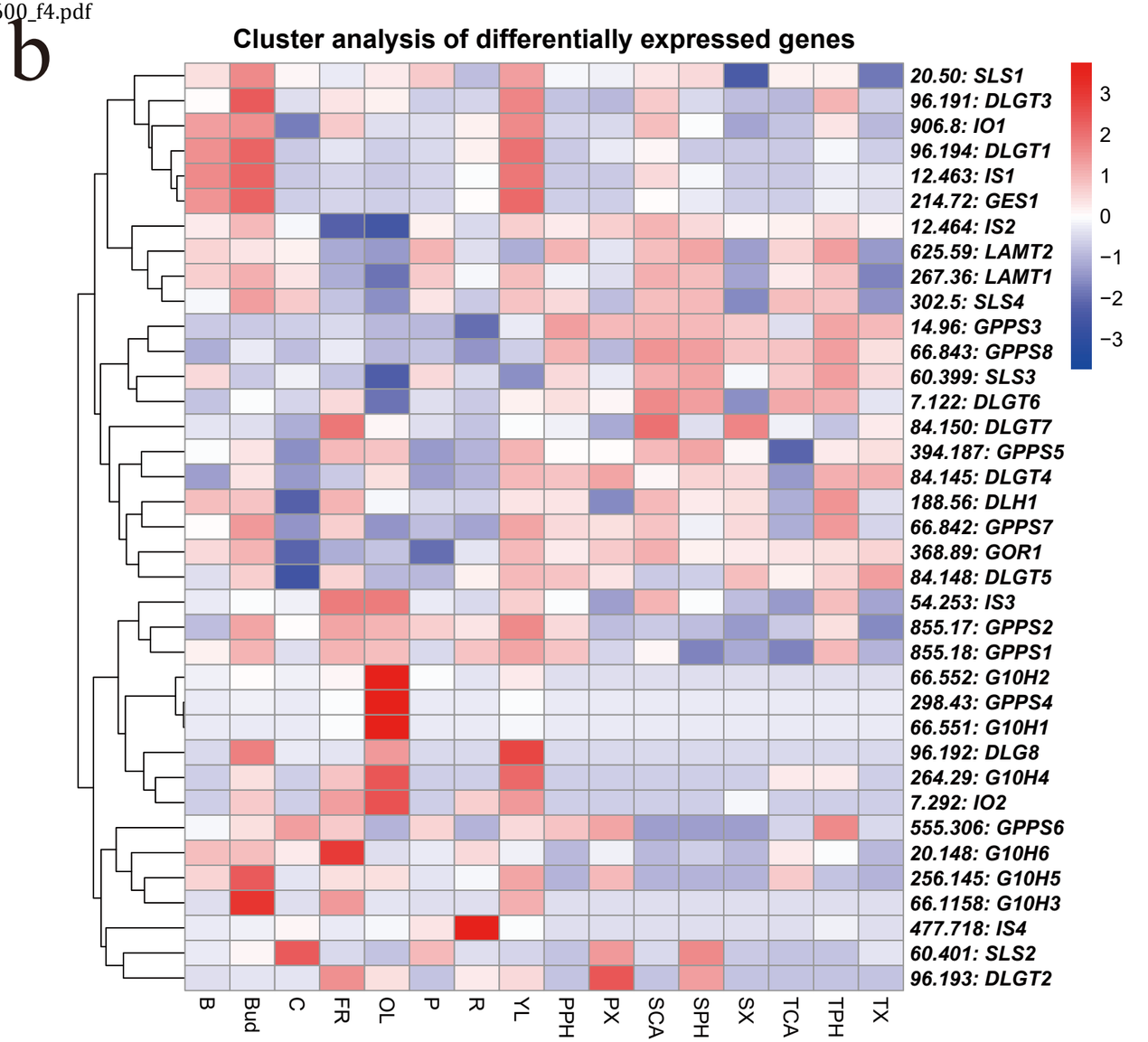
b

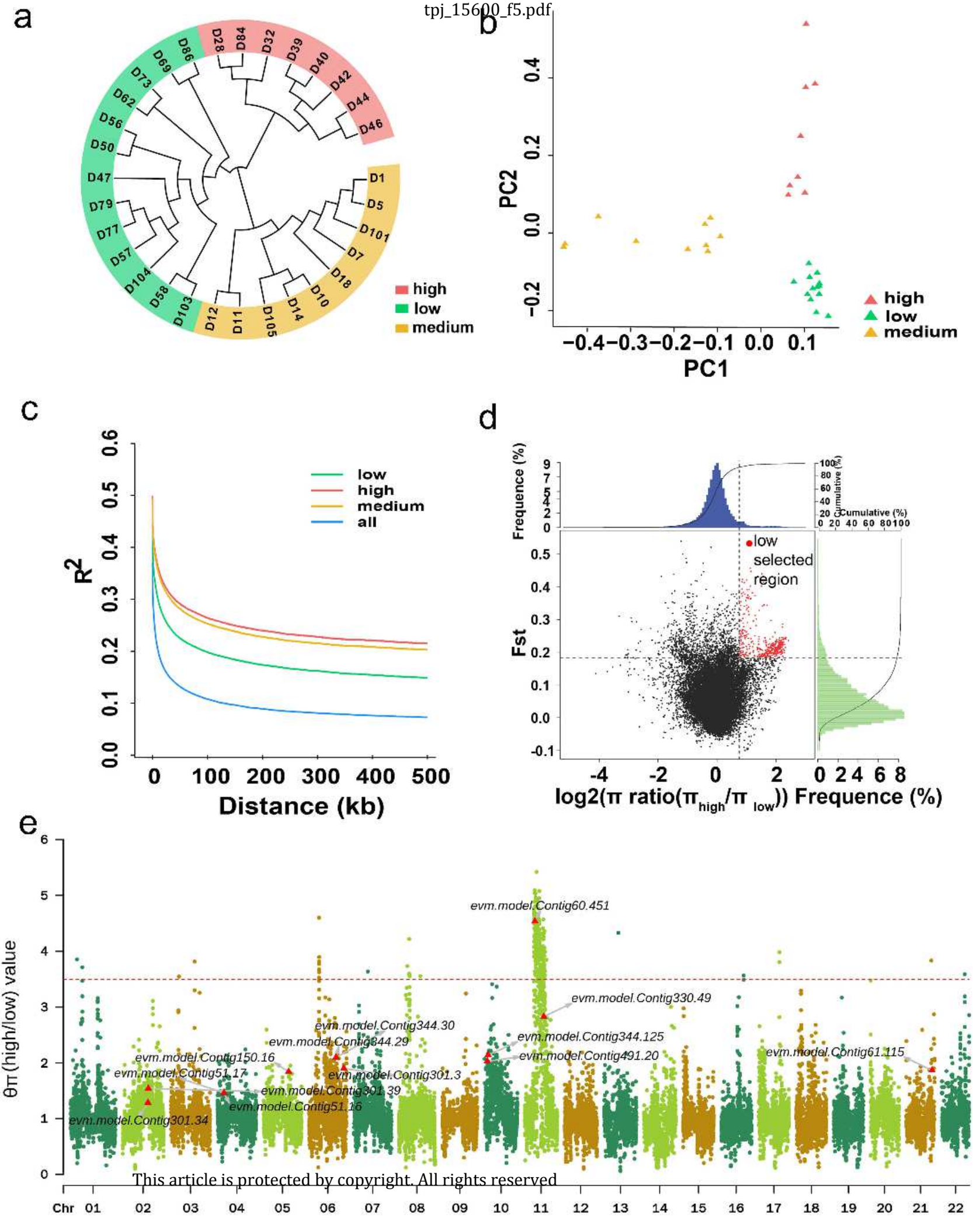


c









a

NcSMASH35
NcMCL22
NcMCL18
NcSMASH9
NcMCL20
NcSMASH65
NcMCL7
NcMCL9

STR CYP CPS MFS TPS ACT RPS
BGLU GES CHX SLS MATE TDC other
10 kb

b

C. canephora chr5
23.19-24.87Mb
23.31-25.13Mb
N. cadamba chr10
3.78-5.73Mb
chr22

evm.model.Contig708.58 evm.model.Contig708.57 evm.model.Contig708.56 evm.model.Contig708.54 evm.model.Contig708.53 evm.model.Contig708.52 evm.model.Contig708.50 evm.model.Contig267.36 evm.model.Contig267.33 evm.model.Contig188.31 evm.model.Contig188.32 evm.model.Contig188.42 evm.model.Contig188.43 evm.model.Contig188.56 evm.model.Contig188.57 evm.model.Contig188.58 evm.model.Contig208.76 evm.model.Contig208.75 evm.model.Contig208.74 evm.model.Contig208.72 evm.model.Contig625.76 evm.model.Contig625.73 evm.model.Contig625.70 evm.model.Contig625.69 evm.model.Contig625.66 evm.model.Contig625.59 evm.model.Contig625.58 evm.model.Contig625.57 evm.model.Contig625.56 evm.model.Contig625.51 evm.model.Contig625.35 evm.model.Contig87.26 evm.model.Contig87.25 evm.model.Contig87.23

B
Bud
C
FR
OL
P
R
YL
PPH
PX
SCA
SPH
SX
TCA
TPH
TX

c

O. pumila chr02
34.31-33.91Mb
N. cadamba chr04
24.15-24.54Mb
C. canephora chr2
47.89-47.25Mb
chr08
27.60-27.51Mb

NcSEH6
MCTS
SWEET17
evm.model.Contig612.
BHLH25
ATX

d

O. pumila chr10
4.82-4.17Mb
N. cadamba chr20
22.00-22.39Mb
C. canephora chr10
6.78-6.58Mb
chr10
4.66-4.17Mb

NcSQE1
evm.model.Contig12.421
BHLH74
ERF
GRF
BHLH85
IS
ISY
NRAMP
CYP81
CYP81