Plant Communications
Research article

CellPress
Partner Journal

Comparative genomics reveals the diversification of triterpenoid biosynthesis and origin of ocotillol-type triterpenes in *Panax*

Zijiang Yang[1,2,6], Xiaobo Li[1,2,6], Ling Yang[1,3,6], Sufang Peng[1,2], Wanling Song[1,2], Yuan Lin[1,2], Guisheng Xiang[1,2], Ying Li[1,2], Shuang Ye[1,2], Chunhua Ma[1,2], Jianhua Miao[4], Guanghui Zhang[1,2], Wei Chen[1,4,5], Shengchao Yang[1,2,*] and Yang Dong[1,4,5,*]

[1]National & Local Joint Engineering Research Center on Germplasm Innovation & Utilization of Chinese Medicinal Materials in Southwest China, Yunnan Agricultural University, Kunming, China

[2]The Key Laboratory of Medicinal Plant Biology of Yunnan Province, Yunnan Agricultural University, Kunming, China

[3]College of Food Science and Technology, Yunnan Agricultural University, Kunming, China

[4]Guangxi Key Laboratory of Medicinal Resources Protection and Genetic Improvement, Guangxi Botanical Garden of Medicinal Plants, Nanning, China

[5]Yunnan Plateau Characteristic Agriculture Industry Research Institute, Kunming, China

[6]These authors contributed equally to this article.

***Correspondence: Shengchao Yang (shengchaoyang@163.com), Yang Dong (dongyang@dongyang-lab.org)**

https://doi.org/10.1016/j.xplc.2023.100591

ABSTRACT

Gene duplication is assumed to be the major force driving the evolution of metabolite biosynthesis in plants. Freed from functional burdens, duplicated genes can mutate toward novelties until fixed due to selective fitness. However, the extent to which this mechanism has driven the diversification of metabolite biosynthesis remains to be tested. Here we performed comparative genomics analysis and functional characterization to evaluate the impact of gene duplication on the evolution of triterpenoid biosynthesis using *Panax* species as models. We found that whole-genome duplications (WGDs) occurred independently in Araliaceae and Apiaceae lineages. Comparative genomics revealed the evolutionary trajectories of triterpenoid biosynthesis in plants, which was mainly promoted by WGDs and tandem duplication. Lanosterol synthase (LAS) was likely derived from a tandem duplicate of cycloartenol synthase that predated the emergence of Nymphaeales. Under episodic diversifying selection, the LAS gene duplicates produced by γ whole-genome triplication have given rise to triterpene biosynthesis in core eudicots through neofunctionalization. Moreover, functional characterization revealed that oxidosqualene cyclases (OSCs) responsible for synthesizing dammarane-type triterpenes in *Panax* species were also capable of producing ocotillol-type triterpenes. Genomic and biochemical evidence suggested that *Panax* genes encoding the above OSCs originated from the specialization of one OSC gene duplicate produced from a recent WGD shared by Araliaceae (Pg-β). Our results reveal the crucial role of gene duplication in diversification of triterpenoid biosynthesis in plants and provide insight into the origin of ocotillol-type triterpenes in *Panax* species.

Keywords: gene duplication, *Panax* genomes, whole-genome duplications, triterpenoid biosynthesis, ocotillol-type triterpenes, specialization

Yang Z., Li X., Yang L., Peng S., Song W., Lin Y., Xiang G., Li Y., Ye S., Ma C., Miao J., Zhang G., Chen W., Yang S., and Dong Y. (2023). Comparative genomics reveals the diversification of triterpenoid biosynthesis and origin of ocotillol-type triterpenes in *Panax*. Plant Comm. **4**, 100591.

INTRODUCTION

Plants have evolved to synthesize a diverse array of metabolites that play essential roles in various biological processes. The adaptivity derived from these metabolites has driven the evolution of plants and even their interactors. For decades, biologists have been intrigued by the evolutionary mechanism underlying

Published by the Plant Communications Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.

the diversification of metabolite biosynthesis in the plant kingdom. Gene duplication is proposed to be the major force driving the evolution of metabolite biosynthesis: relaxed from functional constraints, one duplicate can accumulate mutations. In most cases, such mutations will result in gene loss, but some may be fixed owing to selective advantages conferred by their altered function, whether neofunctionalization, subfunctionalization, or specialization (Ober, 2005). These novelties in function or expression pattern would gradually reshape the biosynthetic pathway for metabolites. In land plants, pervasive whole-genome duplications (WGDs) or polyploidizations serve as the primary sources of gene duplicates. These frequent WGDs are thought to have a key causal role in species diversification, phenotypic evolution, and chemical diversification in both gymnosperm and angiosperm lineages (Landis et al., 2018; Lichman et al., 2020; Stull et al., 2021). The causal linkage between WGDs and diversification of metabolite biosynthesis, although supported on a theoretical basis, remains to be rigorously tested.

Triterpenoids are one of the most diverse metabolites present in plants. Their biosynthesis is catalyzed by enzymes known as oxidosqualene cyclases (OSCs), which can cyclize the precursors 2,3-oxidosqualene and 2,3; 22,23-dioxidosqualene. Two different types of substrate conformation exist during the cyclization process: the chair-boat-chair (CBC) conformation and the chair-chair-chair (CCC) conformation (Thimmappa et al., 2014). Sterols, including cycloartenol and lanosterol, are produced via CBC folding, whereas triterpenes are produced via CCC folding. Based on the catalytic products, plant OSCs can be broadly classified into cycloartenol synthase (CAS), lanosterol synthase (LAS), lupeol synthase (LUS), β-amyrin synthase, and other multifunctional triterpene synthases (bAS and other mTTSs). Sterols function as important membrane components and also as plant hormones that regulate growth and development (Schaller, 2003). The "nonessential" triterpenes are considered to have more specialized functions in plant defense and microbiome interactions (Delis et al., 2011; Khakimov et al., 2015; Miettinen et al., 2018; Huang et al., 2019; Li et al., 2021b; Busta et al., 2021). Genomic screening of the Viridiplantae phylogeny revealed that angiosperms are a hotspot of OSC diversification. Both divergent and convergent evolutionary processes are thought to have influenced the evolution of OSCs, and it is generally accepted that expansion of OSCs has been driven mainly by tandem duplications and that the triterpene synthases of eudicots likely originated from LAS rather than CAS (Pichersky and Lewinsohn, 2011; Xue et al., 2012; Zhou et al., 2016; Cárdenas et al., 2019; Dong et al., 2021). However, the impact of WGDs on the diversification of OSCs and the corresponding evolutionary trajectory remain unresolved.

The genus *Panax* L. (Araliaceae), which contains seven well-defined species and one species complex, is one of the most medicinally important plant genera. The pharmaceutical activities of *Panax* species have been attributed mainly to ginsenosides (glycosylated triterpenoids) (Leung and Wong, 2010; Fan et al., 2020). Biochemical approaches have revealed a wide variety of triterpenoids in *Panax* species, including the dammarane, α/β-amyrin, and ocotillol types (Hou et al., 2021). To date, OSC genes responsible for synthesis of dammarane-type triterpenes have been reported for several *Panax* species (Tansakul et al., 2006; Wang et al., 2014), but the biosynthetic pathway of ocotillol-type triterpenes remains unclear. As one *Panax* species with high medicinal value, *Panax vietnamensis* var. *fuscidiscus* is widely cultivated in Yunnan, China. The high content of ocotillol-type saponins in *P. vietnamensis* var. *fuscidiscus* make it a suitable model for exploring the mechanism of ocotillol-type triterpene biosynthesis (Zhang et al., 2015). *Panax* species have experienced several rounds of WGD in their evolutionary history, but whether extra WGDs have occurred in the common ancestor of all Apiales species after the γ whole-genome triplication (WGT) remains a topic of controversy (Kim et al., 2018a; Li et al., 2021a; Yang et al., 2021a, 2021b; Song et al., 2021). Regardless of disputes about WGD history, genomic and phytochemical evidence indicates that the evolution of triterpenoid biosynthesis in *Panax* species is likely to have been affected by WGDs (Li et al., 2021a). The diversity of triterpenoids and the presence of WGDs in *Panax* species make this genus a suitable model for examining the effects of WGDs on the evolution and diversification of OSCs.

Here we report a high-quality chromosome-level assembly for *P. vietnamensis* var. *fuscidiscus*, together with an improved assembly for *Panax notoginseng*. We found that WGDs have occurred independently in Araliaceae and Apiaceae species rather than being shared by Apiales. Comparative genomics revealed that the diversification of triterpenoid biosynthesis was promoted mainly by WGDs and tandem duplications. Notably, the dammarenediol-II synthases (DDSs) in *Panax* species were functionally characterized as mTTSs. These *Panax* DDS genes originated from the specialization of one OSC gene duplicate produced by the Pg-β WGD. Our findings systematically reveal how gene duplication drives the diversification of triterpenoid biosynthesis in plants and reveal the origin of ocotillol-type triterpenes in *Panax* species.

RESULTS

Panax genome sequencing, assembly, and annotation

PacBio long reads were used to build a *de novo* assembly for *P. vietnamensis* var. *fuscidiscus* (Supplemental Figure 1A). This preliminary assembly was polished with Illumina short reads and scaffolded using Hi-C technology. The final chromosome-level assembly of *P. vietnamensis* var. *fuscidiscus* spans 1.73 Gb, with a scaffold N50 of 144.08 Mb (Supplemental Table 1). The largest 12 scaffolds, which correspond to the karyotype of *P. vietnamensis* var. *fuscidiscus* (2n = 2x = 24), covered 91.04% of the assembly (1.57 Gb) (Supplemental Figures 1B, 2A). The size of the pseudochromosomes is close to the flow cytometry (1.61 Gb) and k-mer-based estimates (1.43 Gb) (Supplemental Figure 3A; Supplemental Tables 2, 3). To evaluate the quality of the *P. vietnamensis* var. *fuscidiscus* genome, 229.47 Gb of the Illumina short reads (132.64×) were mapped to the assembly. The mapping rate of properly paired reads and genome coverage rate were 94.47% and 97.40%, respectively (Supplemental Table 4). We annotated 36 454 protein-coding genes in the *P. vietnamensis* var. *fuscidiscus* genome, with an average gene length of 6166.47 bp (Supplemental Table 5). A total of 33 570 (92.09%) predicted genes could be functionally annotated (Supplemental Table 6). Benchmarking Universal Single-Copy Orthologs (BUSCO) completeness of the assembly and annotated genes were 95.3% and 92.6% (Supplemental

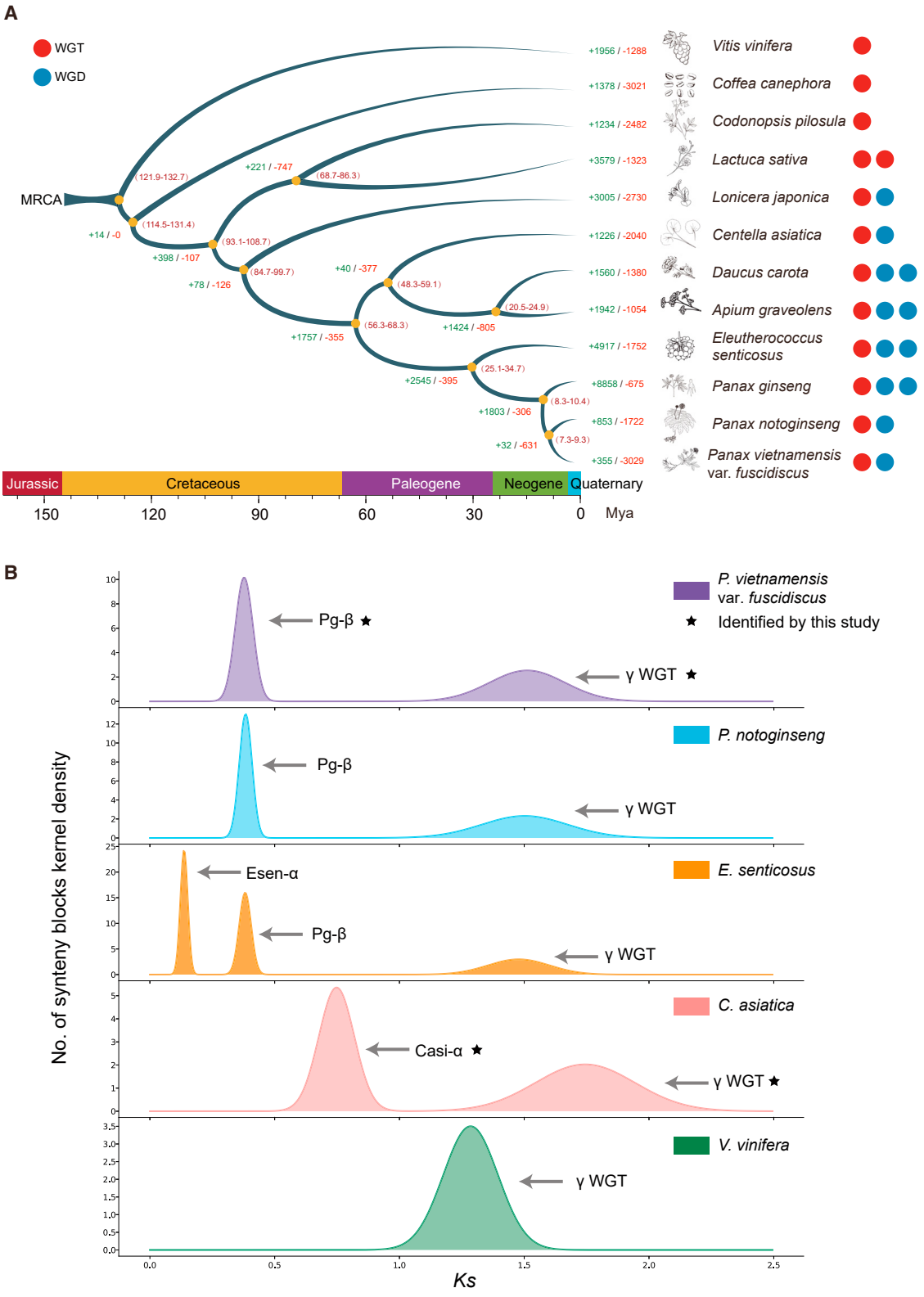

Figure 1. Evolutionary analysis of *P. vietnamensis* var. *fuscidiscus*.

(A) Species tree for 12 eudicots including *P. vietnamensis* var. *fuscidiscus*. Numbers in parentheses indicate estimated divergence times in Mya with 95% confidence intervals. Expansion and contraction of gene families are indicated with plus and minus signs. Whole-genome duplications (WGDs) and whole-genome triplications (WGTs) in each species are shown in blue/red circles. MRCA, most recent common ancestor.

(B) *Ks* distribution of intraspecific collinear blocks. *Ks* peaks of polyploidizations are labeled for each species. Esen, *E. senticosus*; Casi, *C. asiatica*.

Figure 3B; Supplemental Table 7). The reported genome assembly size for *P. vietnamensis* Ha *et* Grushv. (2n = 2x = 24) is 3.00 Gb (Tien et al., 2021), which is 1.73-fold larger than the genome size of *P. vietnamensis* var. *fuscidiscus*. Thus, *P. vietnamensis* var. *fuscidiscus* is likely to be an independent species rather than a variety of *P. vietnamensis* Ha *et* Grushv. If not, *P. vietnamensis* var. *fuscidiscus* is still worth studying for its different genome size compared with *P. vietnamensis* Ha *et* Grushv. and well-developed biosynthetic modules based on *P. vietnamensis* var. *fuscidiscus*.

We also provide an improved chromosome-level assembly of *P. notoginseng* (2n = 2x = 24) created using a previous contig-level assembly (Supplemental Table 8). More sequences (94.29%) were anchored to the 12 pseudochromosomes compared with the previous assembly (86.87%) (Supplemental Figure 2B). We annotated 36 747 protein-coding genes in the updated *P. notoginseng* genome, 92.79% of which were functionally annotated (Supplemental Tables 5 and 9). BUSCO analysis suggested 97.5% and 93.3% completeness of the updated *P. notoginseng* assembly and annotated genes (Supplemental Figure 3B; Supplemental Table 7).

Species-specific LTR expansion produced genome size variation in *Panax*

Repetitive elements constitute 86.79% and 88.18% of the *P. vietnamensis* var. *fuscidiscus* and *P. notoginseng* assemblies. LTR-RTs are the most abundant type of transposable elements (TEs) in both *Panax species*, accounting for 78.94% and 80.66% of the *P. vietnamensis* var. *fuscidiscus* and *P. notoginseng* assemblies. Among the LTR-RTs in *P. vietnamensis* var. *fuscidiscus*, Gypsy elements (54.52% of the genome) are far more abundant than Copia elements (5.67%). A similar phenomenon was observed in the *P. notoginseng* genome, with Gypsy and Copia elements accounting for 55.49% and 4.55% of the genome. DNA transposons are the second most abundant type of TE and constitute 2.90% and 3.13% of the *P. vietnamensis* var. *fuscidiscus* and *P. notoginseng* genomes (Supplemental Tables 10 and 11). Based on intact LTR-RTs (21 251 in *P. vietnamensis* var. *fuscidiscus* and 24 899 in *P. notoginseng*), we estimated the insertion times for LTR-RTs. *P. vietnamensis* var. *fuscidiscus* was found to have experienced a more recent burst of LTRs compared with *P. notoginseng* (Supplemental Figure 4). Clade-level classification of TEs revealed that the numbers of several Gypsy-clade elements (mainly Tekay, Ogre, and Athila) are much higher in *P. notoginseng* than in *P. vietnamensis* var. *fuscidiscus* (Supplemental Tables 12 and 13). These results indicate that *P. notoginseng* experienced a more intense expansion of LTR insertions compared with *P. vietnamensis* var. *fuscidiscus* after their divergence; the difference in genome sizes between the two *Panax* species (~680 Mb) can be attributed mainly to the more intense expansion of LTRs in *P. notoginseng* (LTR size difference, ~609 Mb).

Phylogenomics and evolution of *P. vietnamensis* var. *fuscidiscus*

To study the evolutionary history of *Panax* species, we first performed gene family analysis using 12 eudicots: *Vitis vinifera*, *Coffea canephora*, *Codonopsis pilosula*, *Lactuca sativa*, *Lonicera japonica*, *Centella asiatica*, *Daucus carota*, *Apium graveolens*, *Eleutherococcus senticosus*, *Panax ginseng*, *P. notoginseng*, and *P. vietnamensis* var. *fuscidiscus*. A total of 30 074 ortholog groups, harboring 93.1% of all the studied genes, were identified for the 12 species, and 168 groups are presented as single-copy orthogroups (Supplemental Figure 5A; Supplemental Table 14). Investigation of gene families in *P. vietnamensis* var. *fuscidiscus* and five other Apiales species suggested that *P. vietnamensis* var. *fuscidiscus* and *P. notoginseng* contain 436 and 673 unique gene families, respectively (Supplemental Figure 5B).

Single-copy orthogroups were used to construct maximum likelihood (ML) phylogenetic trees. The species trees inferred by the concatenation method and coalescence-based phylogenetic analysis are identical and well supported (Supplemental Figure 6A and 6B). *P. vietnamensis* var. *fuscidiscus* is placed as a sister lineage to *P. notoginseng* rather than *P. ginseng*, consistent with a *Panax* phylogeny based on chloroplast genomes and ribosomal DNA (Ji et al., 2019). Divergence times were estimated using MCMCTree with time calibrations. The estimated divergence between Araliaceae and Apiaceae occurred ~56.3–68.3 million years ago (Mya). In the *Panax* genus, the speciation of *P. ginseng* occurred first (~8.3–10.4 Mya), followed by the divergence of *P. vietnamensis* var. *fuscidiscus* and *P. notoginseng* (7.3–9.3 Mya) (Figure 1A). We also noted the early divergence of *C. asiatica* in the family Apiaceae, which occurred approximately 48.3–59.1 Mya, validating the basal group position of *C. asiatica* in Apiaceae (Li et al., 2020).

Finally, we estimated the expansion and contraction of gene families during the phylogenetic history of the 12 species using the resolved species tree. In *P. vietnamensis* var. *fuscidiscus*, 355 gene families had undergone expansion, whereas 3029 gene families had undergone contraction ($P < 0.05$) (Figure 1A). Expanded gene families in *P. vietnamensis* var. *fuscidiscus* showed functional enrichment in sesquiterpenoid and triterpenoid biosynthesis ($P < 0.05$) (Supplemental Figure 7A; 7B, Supplemental Tables 15 and 16).

Polyploidization history in Apiales

Polyploidizations in Apiales were systematically characterized to study their impact on the evolution of triterpenoid biosynthesis. The *V. vinifera* genome was used as an outgroup because only one polyploidization event (γ WGT) occurred during its evolution. We inferred WGDs and speciation events by examining the synonymous substitutions per synonymous site (*Ks*) of collinear gene pairs and intra/interspecific syntenic relationships. Two clear peaks were observed in the *Ks* distribution of intraspecific collinear gene pairs for *P. vietnamensis* var. *fuscidiscus*, suggesting an extra round of WGD after the γ WGT (Figure 1B). Interspecific synteny comparison between genomes of *P. vietnamensis* var. *fuscidiscus* and *V. vinifera* revealed that for each genomic region in *V. vinifera*, there are up to six syntenic matches in *P. vietnamensis* var. *fuscidiscus*, validating the extra round of WGD in the latter species (Supplemental Figure 8). In addition to recent peaks attributed to speciation, extra peaks were detected in the *Ks* distribution of collinear gene pairs between *P. vietnamensis* var. *fuscidiscus* and the other two Araliaceae species (*P. notoginseng* and *E. senticosus*) (Supplemental Figure 9). The ancient peaks indicate the shared γ WGT, and the relatively young peaks may

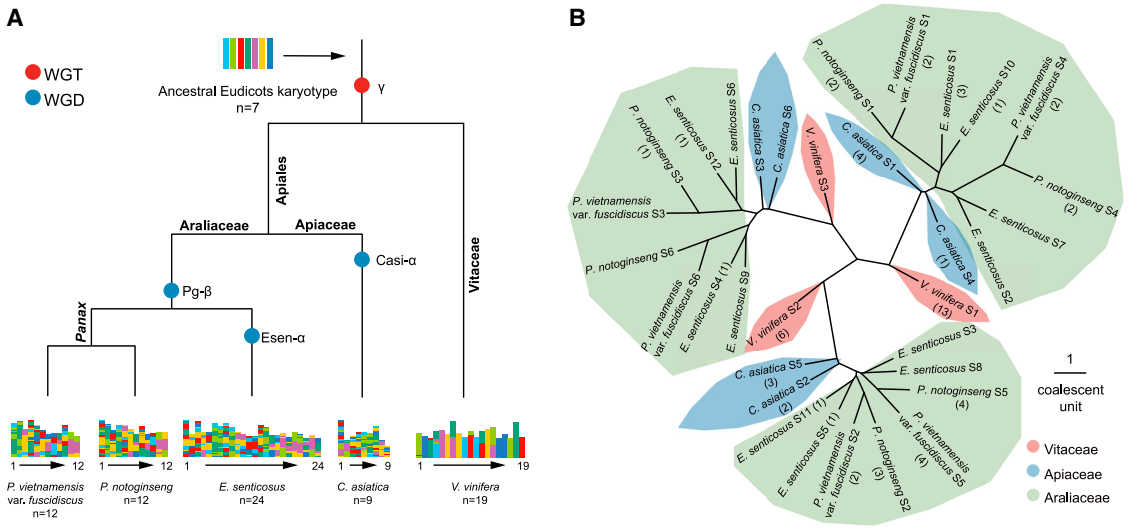

Figure 2. Inference of polyploidization and speciation history in Apiales.

(A) The inferred phylogeny of Apiales species with placement of polyploidizations. Karyotypes were painted in seven colors, corresponding to the seven ancestral eudicot chromosomes.

(B) Synteny-based coalescent species tree showing independent WGDs in Apiaceae and Araliaceae. Genomes were classified into collinear subsets based on polyploidization history (denoted with S). Branch lengths are shown in coalescent units. Because the ASTRAL tree leaves the branch length of terminal branches empty, the lengths of terminal branches were all set to one. Numbers in parentheses represent the number of putative OSC genes.

represent the Pg-β WGD, which is presumed to be shared by Araliaceae species. The *Ks* peak values for the Pg-β WGD and the γ WGT are nearly identical in the three Araliaceae species, suggesting little variation in evolutionary rates among Araliaceae. Synteny comparisons of the updated *P. notoginseng* assembly with that of *E. senticosus* showed exactly 1:2 ratios for the best-matched regions in the largest 12 pseudochromosomes, validating the high quality of the updated *P. notoginseng* assembly compared with an older version (Supplemental Figure 10A and 10B). *C. asiatica* experienced two WGDs according to our analysis (Figure 1B). The *Ks* distribution of interspecific collinear gene pairs between *C. asiatica* and *P. vietnamensis* var. *fuscidiscus* showed two peaks, which correspond to speciation (~0.53) and the shared γ WGT (~1.63) (Supplemental Table 17). The absence of additional peaks suggested that the younger WGDs in Apiaceae and Araliaceae may have occurred independently after their speciation (Figure 2A).

To examine Apiales evolution with greater resolution, we performed synteny-based phylogenetic analysis. Five species (*V. vinifera*, *C. asiatica*, *E. senticosus*, *P. notoginseng*, and *P. vietnamensis* var. *fuscidiscus*) that exhibit a well-preserved ancestral eudicot karyotype (AEK) were selected for the analysis. Using the AEK and the *V. vinifera* genome as references, collinear regions were partitioned into different copies for each species with consideration of WGDs (Supplemental Figures 11–15; Supplemental Table 18). Based on 2255 collinear gene pairs (23 821 genes), ASTRAL produced a phylogenetic tree for the five species with a normalized quartet score of 0.8146. The topology of the synteny-based species tree provides solid evidence that the Pg-β WGD occurred independently in Araliaceae and was shared by Araliaceae species (Figure 2B and Supplemental Figure 16). Interestingly, the collinear subsets for *C. asiatica* in all three lineages produced by the γ WGT did not form sister groups but split successively instead. This suggested that the relatively recent WGD in *C. asiatica* may have been induced by an ancient hybridization.

Evolution of OSCs was mainly promoted by WGDs and tandem duplications

Previous studies have suggested that OSCs for sterol biosynthesis in Eukarya have a bacterial origin and that plant OSCs have likely undergone divergent evolution, with triterpene biosynthesis derived from sterol biosynthesis (Xue et al., 2012; Santana-Molina et al., 2020). Thus CAS likely served as the foundation of OSC evolution. Here, we performed phylogenetic and comparative genomics analyses to clarify the evolution of OSCs in plants. Nine species (*Amborella trichopoda*, *Aristolochia fimbriata*, *V. vinifera*, *C. asiatica*, *E. senticosus*, *P. ginseng*, *P. vietnamensis* var. *fuscidiscus*, *P. notoginseng*, and *Panax quinquefolius*) were included in the analysis, including six Apiales species selected for their diversity in triterpenoid biosynthesis and well-characterized phylogenetic history. We included *A. trichopoda* (ANA-grade) and *A. fimbriata* (Magnoliids) in the analysis for their absence of WGD since the emergence of flowering plants (Amborella Genome Project, 2013; Qin et al., 2021). First, we performed genome-wide identification of OSCs based on conserved protein domains. For *P. quinquefolius*, which lacks a reference assembly, one DDS was used (Supplemental Table 19). In contrast to the abundant OSC genes in eudicots, we identified only one putative OSC in *A. trichopoda* and two putative OSCs in *A. fimbriata*, implying an important role for WGDs in the expansion of OSCs. An ML phylogenetic tree was built for the identified putative OSCs using codon alignments (Figure 4A). On the basis of conserved motifs (Supplemental Figure 17) and phylogenetic relationships with functionally characterized OSCs, the OSCs were classified into putative functional groups (CAS, LAS, LUS, and bAS and other mTTSs)

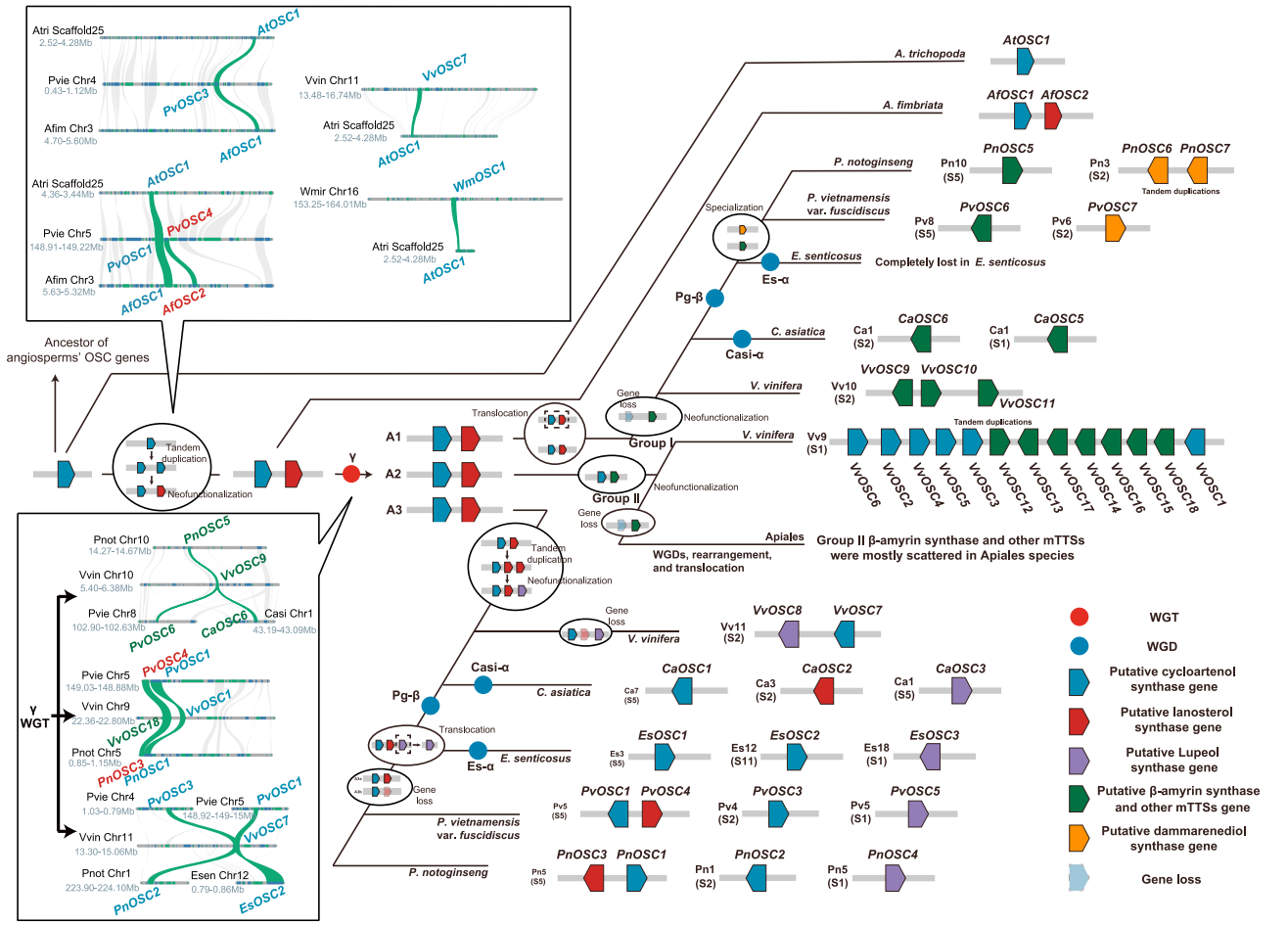

Figure 3. The inferred evolutionary trajectory of OSC genes in plants.

Polyploidizations are shown with blue and red circles. Pentagons with a solid border represent OSC genes. Gene loss is shown by pentagons with a dashed border. Interspecific micro-syntenic relationships of putative OSC genes are shown in boxes. Direct collinear relationships between putative OSC genes are highlighted in green (Vvin, *V. vinifera*; Pvie, *P. vietnamensis* var. *fuscidiscus*; Pnot, *P. notoginseng*; Atri, *A. trichopoda*; Afim, *A. fimbriata*; Wmir, *W. mirabilis*). Inferred gene duplication, neofunctionalization, translocation, and loss are highlighted in circles.

(Chen et al., 2021a). Functionally characterized DDSs from *Panax* species were nested within bAS and other mTTSs, indicating their close phylogenetic relationship (Tansakul et al., 2006; Wang et al., 2014). We also noticed that members of bAS and other mTTSs were recovered in two lineages (group I and group II) with high support, suggesting their distinct origins.

The distribution pattern of OSCs on the synteny-based species tree suggested that the expansion of OSCs was affected by WGDs (Figure 2B). In addition, the scarcity of OSCs on the lineage leading to *V. vinifera* S3 indicated that gene loss or translocation events had occurred before speciation. We next performed inter/intraspecific synteny analysis to investigate the evolutionary trajectory of OSCs with differentiation of paralogous and orthologous syntenic regions produced by polyploidizations and speciation. Intraspecific synteny comparisons revealed OSCs produced from recent WGDs (Casi-α and Pg-β WGD) in Apiales species (Supplemental Figure 18). In *P. vietnamensis* var. *fuscidiscus*, no direct syntenic relationship was found for *PvOSC7* (DDS gene) and *PvOSC6* (bAS and other mTTSs), yet both genes were located on highly syntenic chromosomal regions produced by the Pg-β WGD. The same phenomena were also observed in *P. notoginseng* between *PnOSC5* (bAS and other mTTSs) and the tandemly-duplicated *PnOSC6*/*PnOSC7* (DDS genes) (Supplemental Figure 18). Thus, we speculate that DDS in *Panax* species likely originated from neofunctionalization of a group I bAS and other mTTS copy produced from the Pg-β WGD. We observed only one syntenic relationship between *VvOSC12* (from chromosome [chr] 9) and *VvOSC9* (from chr 10) in *V. vinifera*. Considering the fact that OSCs in grape are found on only three chromosomes (chr 9, 10, 11) and that chr 9, chr 11, and a part of chr 4 were produced by the γ WGT, grape OSCs from chr 9, 10, and 11 are likely to share the same origin. After the γ WGT, a chromosomal region harboring OSCs in chr 4 was translocated to chr 10. This assumption was verified by interspecific synteny comparisons between *V. vinifera* and Apiales species, in which the grape CASs *VvOSC7* (from chr 11) and *VvOSC1* (from chr 9) showed syntenic relationships with the same *P. vietnamensis* var. *fuscidiscus* CAS, *PvOSC1* (Figure 3). We also compared OSC syntenic relationships of *P. vietnamensis* var. *fuscidiscus* and *V. vinifera* with species with non-duplicated genomes (*A. trichopoda* and *A. fimbriata*). The CASs produced by the Pg-β WGD (*PvOSC3* and *PvOSC1*)

and an LAS (*PvOSC4*) from *P. vietnamensis* var. *fuscidiscus* showed clear syntenic relationships with OSC genes from *A. trichopoda* (*AtOSC1*) and *A. fimbriata* (*AfOSC1* and *AfOSC2*). A syntenic relationship was also found between grape *VvOSC7* and *A. trichopoda AtOSC1* (Figure 3). Such conservation was even detected between *A. trichopoda* and the gymnosperm *Welwitschia mirabilis* (Figure 3), demonstrating that CAS genes are spatially conserved in higher plants.

With the above information, we deduced the evolutionary trajectory of OSCs (Figure 3). OSCs were conserved for sterol biosynthesis during the early stages of plant evolution, as only CASs were found in the genomes of lower plants (Xue et al., 2012). Following the emergence of angiosperms, one CAS duplicate (possibly produced by tandem duplication) may have diversified to give rise to LAS through neofunctionalization. The absence of LAS in *Amborella* suggests that the duplication probably occurred after *Amborella* speciation. The chromosomal region harboring CAS and LAS was triplicated into three copies by the γ WGT (A1–3). Several changes were inferred for the triplicated copies. Before the speciation of grape and Apiales, A1 experienced translocation followed by functional diversification of LAS to group I bAS and other mTTSs. The newly formed group I bAS and other mTTSs was duplicated by the Pg-β WGD, with one copy then neofunctionalizing to DDS in the ancestor of extant *Panax* species. For A2, neofunctionalization of LAS gave rise to group II bAS and other mTTSs. In the lineage leading to Apiales, CAS was lost, and the group II bAS and other mTTSs was likely affected by reshuffling, resulting in a non-syntenic distribution pattern. By contrast, CAS and group II bAS and other mTTSs were retained and proliferated through tandem duplications in grape. LUS may have been produced by neofunctionalization of a tandemly duplicated copy of LAS in A3 before the speciation of grape and Apiales. In the Apiales lineage, the LUS probably experienced translocation, as no syntenic relationships were found between LUS and other OSCs.

Functional characterization revealed the origin of ocotillol-type triterpenes in *Panax*

To verify the proposed Pg-β origin of DDSs in *Panax* species, we performed functional analysis to determine the catalytic activities of each tested OSC. Nine OSC genes (five from group I bAS and other mTTS clades and four from the DDS clade) were selected for the analysis (Figure 4B). The OSC genes were heterologously expressed in mutant yeast strain GIL77, which was engineered to accumulate the precursor oxidosqualene (Morita et al., 1997). The products were identified through GC–MS and NMR (Supplemental Figures 19–26; Supplemental Tables 20 and 21). Nine products were identified for every OSC from group I bAS and other mTTSs. For PvOSC6, PgOSC9, PnOSC5, and CaOSC5, α-amyrin, β-amyrin, ψ-taraxasterol, and 3-epicabraleadiol were identified as the main products, with trace amounts of δ-amyrin, taraxasterol, dammarenediol-II, ocotillol, and an unidentified product. The product profile of CaOSC6 was slightly different, with an increased proportion of ψ-taraxasterol and a decrease in α-amyrin content (Figures 4B, 4C, and 27; Supplemental Table 22). In a recent study, CaOSC5 was functionally characterized as a multifunctional OSC producing δ-amyrin, α-amyrin, β-amyrin, ψ-taraxasterol, taraxasterol, and an unidentified product in a ratio of 1:67:26:4:1:1 (Kim et al., 2018b). The previously reported catalytic activities of CaOSC5 are highly consistent with our results, except for the weak ability to produce dammarane-type triterpenes (dammarenediol-II, ocotillol, and 3-epicabraleadiol) identified in our study. Surprisingly, ocotillol and 3-epicabraleadiol were also detected in addition to dammarenediol-II as catalytic products of DDSs from *Panax* species (PgOSC11, PqDDS, PvOSC7, and PnOSC6) (Figure 4B and Supplemental Figure 27; Supplemental Table 22). This multifunctional nature of DDSs in *Panax* species has not previously been reported. We also noted that PgOSC11 produces mainly ocotillol, whereas the other DDSs predominantly produce dammarenediol-II. These findings validate our assumption that DDS in *Panax* species originated from a duplicate of a group I bAS and other mTTSs produced from a WGD. After the Pg-β WGD, one copy of group I bAS and other mTTSs retained its original function (similar to homologs in *C. asiatica*), whereas the other copy experienced neofunctionalization. Judging by the catalytic products, this neofunctionalization should be viewed as a specialization process: from a generalist ancestor to a more specialized state.

Selective forces underlying evolution of OSCs

According to their deduced evolutionary trajectory, OSC genes have experienced several rounds of independent neofunctionalization and specialization events. It is expected that the diversifications occurred under various selection pressures. To characterize the selective forces driving the evolution of OSC genes, we performed various branch-specific tests.

Branch-site unrestricted statistical test for episodic diversification (BUSTED) analysis found evidence (likelihood ratio test [LRT], $P < 0.05$) of gene-wide episodic diversifying selection on at least one site on at least one branch in the phylogeny (Supplemental Figure 28). Adaptive branch-site random effects likelihood (aBSREL) and mixed effects model of evolution (MEME) were then used to determine the exact lineages and sites that were under positive selection. With *a priori* knowledge that CAS genes serve as a blueprint for functional diversification of OSCs, CAS lineages were labeled as background in the branch-site analysis. Analysis with aBSREL found evidence of episodic diversifying selection on 20 out of 159 branches in the tested phylogeny (LRT, $P \leq 0.05$), with only four in the CAS and LAS clades and the rest distributed in lineages leading to LUS, bAS and other mTTSs, and DDS (Supplemental Figure 29). The fact that almost all of the CAS genes were under negative or neutral selection could be explained by the importance of their cycloartenol product, which is the precursor for almost all plant sterols and plays an essential role in plant developmental processes (Gas-Pascual et al., 2014). Notably, episodic diversifying selection was detected on internal branches leading to LUS and group I/II bAS and other mTTSs (nodes 3–5), in which the major neofunctionalization of OSCs occurred (Supplemental Figure 29). This indicates that the diversification of sterol biosynthesis toward triterpene biosynthesis in core eudicots was driven by episodic positive selection, possibly due to the better adaptability conferred by the triterpene products. Notably, the specialization of DDS from group I bAS and other mTTSs in *Panax* species was predicted to be under neutral or negative selection. This could be explained by trade-offs during specialization: for an enzyme in the generalist state, specialization toward certain functions requires a decrease

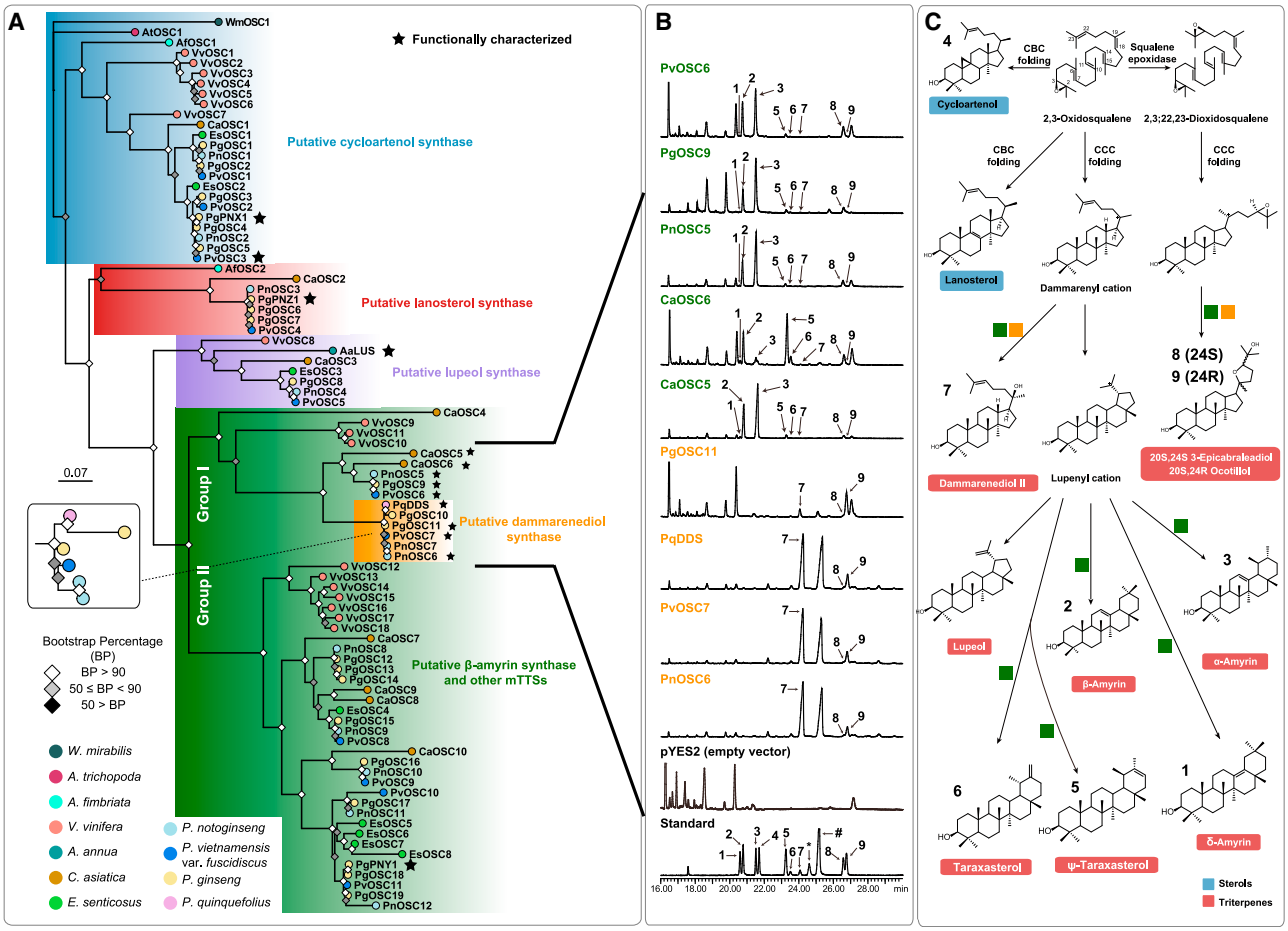

Figure 4. Phylogenetic analysis and functional characterization of OSCs.

(A) ML phylogenetic tree of OSCs based on codon alignments. Bootstraps are shown as colored squares at each node, and species are shown as colored circles at each terminal branch.

(B) Functional characterization of nine OSCs using heterologous expression. The asterisk (*) and hash (#) in the total ion chromatograms (TICs) represent the epoxydammaranes mono-trimethylsilyl ether and dammarenediol-II mono-trimethylsilyl ether, respectively. 1, δ-amyrin; 2, β-amyrin; 3, α-amyrin; 4, cycloartenol; 5, ψ-taraxasterol; 6, taraxasterol; 7, dammarenediol-II; 8, 20S,24S-3-epicabraleadiol; 9, 20S,24R-ocotillol.

(C) Schematic for triterpenoid biosynthesis with sterols highlighted in blue and triterpenes highlighted in red. Compound numbers correspond to the numbers in TICs from **(B)**. Colored squares indicate functions of enzymes in **(B)**.

in the rest. In most cases, reduced promiscuity was shaped by negative selection (Tokuriki et al., 2012; Noda-Garcia and Tawfik, 2020; Tawfik and Gruic-Sovulj, 2020).

MEME found evidence of episodic diversifying selection at 84 sites ($P < 0.05$) (Supplemental Figure 30A). Most of these sites were located near the N/C terminus and the putative active center (Supplemental Figure 30B). Residues from several function-related motifs were found to be under episodic positive selection. For the motif M(W/L)C(Y/H)CR, which has been proposed to stabilize tetracyclic or pentacyclic intermediates, the second site (W/L) was identified as being under episodic positive selection (Kushiro et al., 1999; Ito et al., 2016). Motif Y410, which has been proposed to play an important role in ceiling formation of the active center or in D-ring formation, was also under diversifying selection. The site is conserved as Y in CBC-folding OSCs and F in CCC-folding OSCs (Ma et al., 2016; Chen et al., 2021a). These results provide insights into the evolution of OSCs at the molecular level.

DISCUSSION

The discovered evolutionary trajectories for triterpenoid biosynthesis demonstrate the prominent role of gene duplication in creating a diverse array of triterpenoids in plants. WGDs and tandem duplications are the main forces driving the diversification of OSCs. An ancient tandem duplication of CAS during the early evolution of angiosperms may have given rise to LAS. This event possibly predates the emergence of Nymphaeales species, given the presence of LAS orthologs in *Nymphaea colorata* (Wang et al., 2022). The expansion and diversification of OSCs in core eudicots were attributed mainly to the γ WGT, with subsequent neofunctionalization occurring in the LAS triplicates. Specifically, triterpene synthases in core eudicots originated from independent neofunctionalization of LAS copies. This finding supports the hypothesis that eudicot triterpene biosynthesis

derives from LAS rather than CAS (Kolesnikova et al., 2006; Xue et al., 2012). Indeed, experimental evidence suggested that LAS can supplement the biosynthesis of phytosterols in plants. However, the methyl jasmonate/bacteria-induced regulatory mechanism and tissue-specific expression pattern of LAS suggest its similarity to triterpene synthases (Zimmermann et al., 2004; Kolesnikova et al. 2006). The altered regulation and expression pattern of LAS may represent the initial step toward triterpene biosynthesis. Major angiosperm lineages such as monocots also exhibit great potential for the synthesis of various triterpenes (Inagaki et al., 2011); thus, the revealed γ WGT origin of triterpene biosynthesis in core eudicots suggests convergent evolution in OSC diversification in addition to the prevalent divergent evolution. We also revealed that group I and group II bAS and other mTTSs in eudicots originated from different LAS copies, thus explaining their distant phylogenetic relationship. A recent study of the evolutionary path of OSC genes based on phylogenetic trees inferred several major duplication events for OSC genes, including one ancient duplication event generating the CAS and LAS lineages and another three separate duplication events generating triterpene synthase (LUS and bAS) (Wang et al., 2022). Our results show that the former ancient duplication event was likely caused by tandem duplication of the ancestral OSC gene, whereas the latter three duplication events actually resulted from a single duplication event, the γ WGT. This finding also demonstrates the limitations of using only phylogenetic trees when inferring the evolutionary paths of genes.

Dammarane-type and ocotillol-type triterpenes are abundant mainly in *Panax* species. Several *Panax* DDS genes have been functionally characterized as producing dammarenediol-II, but the genes responsible for synthesizing ocotillol-type triterpenes remain unclear. Our analysis revealed the origin of the DDS gene family in *Panax* species and its multifunctional nature. Future studies using site-directed mutagenesis and crystal structure analysis may provide insight into the reaction mechanism that underlies the shift in product profile of these OSCs.

In principle, reshaping of metabolite biosynthesis after gene duplication is strongly affected by selection. Our results suggest that most of the WGD-derived OSC gene copies were lost during evolution, possibly owing to accumulation of negative mutations under relaxed selection pressure. However, some OSC duplicates acquired altered functions through mutations, which were then fixed by fitting ecological opportunities. This scenario was illustrated by the effect of episodic diversifying selection on neofunctionalization of LAS copies toward triterpene synthases in core eudicots. Such functional innovations of duplicates are not always driven by positive selection. In ancestral lineages of *Panax* species, one group I bAS and mTTS copy (which mainly produced amyrin) that derived from the Pg-β WGD experienced functional specialization under negative/neutral selection, eventually giving rise to DDS. The absence of positive selection might be explained by the trade-off in catalytic activities between amyrin-type and dammarane-type triterpenes (Noda-Garcia and Tawfik, 2020). Accordingly, it should be noted that the absence of positive selection during the creation of novelties does not necessarily indicate a lack of improvement in adaptivity.

In summary, the revealed origin and evolutionary history of triterpenoid biosynthesis in angiosperms provide insight into how gene duplication can drive the diversification of metabolite biosynthesis. In plants, triterpenoids are often further modified by tailoring enzymes (e.g., cytochrome P450s, glycosyltransferases, and acyltransferases) (Thimmappa et al., 2014). Studies have suggested that genes for OSCs and tailoring enzymes are likely functionally co-opted by gene duplication, resulting in diversification in gene regulation and expression patterns for triterpenoid biosynthesis (Li et al., 2021b; Su et al., 2021). Future studies on the interactions between these tailoring enzymes and OSCs and their origins will deepen our understanding of the evolution of metabolite biosynthesis.

METHODS

Genome sequencing and assembly

Plant samples of *P. vietnamensis* var. *fuscidiscus* were collected from individuals cultivated in Jinping County, Yunnan, China. Fresh leaves, stems, and roots were stored in liquid nitrogen and sent to Novogene for sequencing (Beijing, China). High-molecular-weight genomic DNA was extracted from leaves using the cetyltrimethylammonium bromide (CTAB) method and purified with a QIAGEN Genomic Kit (Qiagen, USA). For long-read sequencing, 20-kb SMRTbell libraries were generated and sequenced on the PacBio Sequel platform. This produced ~67.7× PacBio long reads. We also generated ~132.6× Illumina short reads. Four libraries with an insert size of 300 bp were prepared and sequenced on the Illumina HiSeq 2000 platform (Illumina, San Diego, CA, USA). High-throughput chromosome conformation capture (Hi-C) libraries were prepared and sequenced on the Illumina HiSeq 4000 platform. In brief, chromatin was cross-linked with formaldehyde and digested with the restriction enzyme *DpnII* before sequencing. For the purpose of gene prediction, total RNA was isolated from leaves, stems, and roots using the RNAprep Pure Plant Kit (TIANGEN). RNA libraries with an insert size of 300 bp were generated and sequenced on the Illumina HiSeq 2000 platform (Supplemental Table 23).

The genome size of *P. vietnamensis* var. *fuscidiscus* was estimated using flow cytometry (BD FACSCalibur) and GenomeScope (v2.0) with kmer frequencies counted from 132.6× Illumina reads using Jellyfish (v2.2.10) (Marçais and Kingsford, 2011; Ranallo-Benavidez et al., 2020). The PacBio reads were assembled using NextDenovo (v2.4.0) (https://github.com/Nextomics/NextDenovo), followed by two rounds of polishing with NextPolish (v1.3.1) (Hu et al., 2020). After removing allelic contigs using Purge Haplotigs (v1.1.1) (Roach et al., 2018), we performed scaffolding using Juicer (v1.6.2) (Durand et al., 2016) and the three-dimensional (3D) *de novo* assembly (3D-DNA) pipeline (Dudchenko et al., 2017). Mis-joins were manually corrected on the basis of Hi-C contact signals. For transcriptome assembly, raw reads were trimmed with fastp (v0.20.1) (Chen et al., 2018) and assembled using Trinity (v2.11.0) (Grabherr et al., 2011).

The quality of the genome assemblies was evaluated using BUSCO (v5.1.2) (Manni et al., 2021) with dataset eudicots_odb10. We also mapped Illumina reads to the genome using BWA-MEM (v0.7.12) (Li and Durbin, 2009a) and calculated the mapping statistics using SAMtools (v1.9) (Li et al., 2009b).

Genome annotation

We used LTR_FINDER_parallel (v1.1) (Ou and Jiang, 2019) and LTRharvest (v1.0) (Ellinghaus et al., 2008) to predict long terminal repeat retrotransposons (LTR-RTs). The identified LTR-RT candidates were passed to LTR_retriever (v2.8) (Ou and Jiang, 2018b) to filter out the false positives and generate a genome LTR assembly index (LAI) (Ou et al., 2018a). Only intact LTRs were retained for insertion time estimation. The equation $T = K/2\mu$ was used for time estimation, where K is the LTR divergence rate and μ is the neutral mutation rate (1.3×10^{-8} mutations per site per year). We also used RepeatModeler (v2.0) (Flynn et al., 2020) to detect novel repeat sequences. Repetitive elements generated by LTR_retriever and RepeatModeler were fed to RepeatMasker (v4.0.9) (http://www.repeatmasker.org) for *de novo* prediction. For evidence-based methods, repetitive elements were predicted using RepeatMasker and RepeatProteinMask (v4.0.9) (http://www.repeatmasker.org) with Repbase (v24.06) (Bao et al., 2015) as the reference. The tandem repeats were annotated using Tandem Repeat Finder (v4.09) (Benson, 1999). The predicted LTR-RTs were further classified by TEsorter (v1.2.5) (Zhang et al., 2022) with REXdb Viridiplantae (v2.2) (Neumann et al., 2019).

Gene structures were predicted using a combination of *ab initio*-, homology-, and transcript-based methods. GenScan (v1.0) (Aggarwal and Ramaswamy, 2002), GlimmerHMM (v3.0.3) (Majoros et al., 2004), geneid (v1.4.4) (Alioto et al., 2018), Augustus (v3.2.2) (Stanke et al., 2008), and SNAP (v1.0) (Korf, 2004) were used for *ab initio* prediction of protein-coding genes. For the homology-based method, proteomes of *A. thaliana*, *V. vinifera*, *E. senticosus*, *D. carota*, and *P. ginseng* were searched against the genomes using TBLASTN (v2.2.29+) (Altschul et al., 1990) with $1e^{-5}$ as the cutoff e-value. Gene models were predicted by GenomeThreader (v1.7.3) (Gremme et al., 2005) using the above hits. For the transcript-based method, Program to Assemble Spliced Alignments (PASA) (v2.4.1) (Haas et al., 2003) was used for gene prediction by comparing Trinity transcripts with genomes. Finally, all gene models were integrated using EVidenceModeler (v1.1.1) (Haas et al., 2008) and updated with PASA. Functional annotation was performed with eggNOG-mapper (v2.1.7) (Cantalapiedra et al., 2021) by searching the eggNOG database (v5.0.2) (Huerta-Cepas et al., 2019) (Viridiplantae-33090) using DIAMOND (v2.0.14) (Buchfink et al., 2015).

Phylogenomics and evolutionary analysis

Orthogroups were identified using OrthoFinder (v2.5.4) (van Dongen, 2000; Emms and Kelly, 2019) based on protein sequences of 12 species (Supplemental Table 24). The Venn diagram was visualized using Evenn (Chen et al., 2021b).

Species trees were inferred based on single-copy orthogroups. Protein sequences from each single-copy orthogroup of 12 species were extracted and aligned using MAFFT (v7.475) (Katoh and Standley, 2013). Then, the protein alignments were converted to codon alignments using PAL2NAL (v14) (Suyama et al., 2006). Poorly aligned regions from codon alignments were trimmed using trimAl (v2.rev0) (Capella-Gutiérrez et al., 2009). For the concatenation-based method, an ML phylogenetic tree was built based on concatenated codon alignments using IQ-TREE (v2.0.3) (Nguyen et al., 2015) with the best-fit substitution model determined using ModelFinder (Kalyaanamoorthy et al., 2017). Branch supports were estimated using 1000 replicates with ultrafast bootstrap approximation (UFBoot2) (Hoang et al., 2018). For the coalescent-based method, a species tree was estimated using ASTRAL (v5.7.7) (Zhang et al., 2018) based on ML trees produced from IQ-TREE. We estimated species divergence times using MCMCTree from the PAML package (v4.9j) (Yang, 2007) with molecular clock and nucleotide substitution set as correlated rates and JC69 model. The MCMC process was run for 100 000 iterations with a burn-in of 50 000 and a sampling frequency of five. The tree was calibrated with the following constraints: divergence time of *D. carota* and *A. graveolens* (~22–37 Mya), divergence time of Araliaceae and Apiaceae (~45–70 Mya), and divergence time of *V. vinifera* and the other studied species (~111–131 Mya) (Kumar et al., 2017). Phylogenetic trees were visualized using FigTree (v1.4.4) (http://tree.bio.ed.ac.uk/software/figtree/).

Changes in gene family size during species evolution were estimated using CAFE (v5) (Mendes et al., 2020). Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses of gene families were performed using clusterProfiler (v4.2.2) (Wu et al., 2021) and TBtools (v1.098685) (Chen et al., 2020), respectively, with *P* values adjusted by the Benjamini and Hochberg method.

WGD and speciation analysis

The WGDI toolkit (v0.5.1) (Sun et al., 2022) was used to detect WGD and speciation events. First, BLASTP (v2.2.29+) was used to search for homologs with a $1e^{-5}$ cutoff e-value. Collinear genes were identified by WGDI on the basis of the identified homologs using the parameter -icl. *Ks* values of collinear gene pairs were then calculated using the YN00 program from the PAML package with the Nei-Gojobori method (Nei and Gojobori, 1986). The median *Ks* values of inter/intraspecific collinear blocks were fitted using Gaussian kernel density estimation with the parameter -pf. The intraspecific syntenic relationships within *P. vietnamensis* var. *fuscidiscus*, together with the GC content, TE density, and gene density, were visualized using Circos (v0.69-9) (Krzywinski et al., 2009).

Synteny-based phylogenetic analysis was used to infer the WGD and speciation history. On the basis of the similarity and completeness of inter/intraspecific syntenic blocks, syntenic regions were assigned to WGD-related putative sets for *V. vinifera*, *C. asiatica*, *E. senticosus*, *P. notoginseng*, and *P. vietnamensis* var. *fuscidiscus* with parameters -bi and -a. Collinear genes from the characterized sets were extracted and used to construct ML phylogenetic trees separately using IQ-TREE. Collinear gene pairs encompassing genes from all studied species were retained for ASTRAL analysis.

Inferring evolutionary trajectories of OSC genes

Putative OSCs were identified using HMMER (v3.1b2) (Eddy, 1998) by searching with the squalene-hopene cyclase N-terminal domain (PF13249) and C-terminal domain (PF13243) from Pfam (v35.0) (Mistry et al., 2021) with the parameter -cut_tc. Sequences that contained both domains were retained for analysis. For phylogenetic analysis, protein sequences of the putative OSCs

were aligned using MAFFT. The protein alignments were converted to codon alignments by PAL2NAL, followed by trimming with trimAl. IQ-TREE was used to construct an ML phylogenetic tree for the putative OSCs. The tree and motifs were visualized using the R packages ggtree (v2.4.1) (Yu et al., 2018) and ggmsa (v1.0.0) (http://yulab-smu.top/ggmsa/). To assist with classification of putative OSCs, sequences of functionally characterized OSCs were downloaded from NCBI and included in the analysis. One *P. vietnamensis* var. *fuscidiscus* OSC (PvOSC3) was also functionally characterized (Supplemental Figure 31). For synteny-based analysis, the syntenic relationships among putative OSC genes were identified with WGDI. We used JCVI utility libraries (v1.1.23) (Tang et al., 2008) to visualize the micro-synteny of OSCs.

Functional characterization of OSCs

Nucleotide coding sequences of putative OSC genes were synthesized and ligated into the yeast expression vector pYES2 (Invitrogen) under the control of the *GAL1* promotor by GeneCreate (Wuhan, China). Vectors carrying putative OSC genes were transformed into DH5α competent cells. The resulting plasmid DNAs were transformed into the mutant yeast strain GIL77 by the lithium acetate/single-stranded carrier DNA/PEG method (Gietz and Schiestl, 2007). Yeast strains transformed with the empty vector were used as controls. Yeast strains were incubated in synthetic medium containing ergosterol (20 μg ml^{-1}), hemin chloride (13 μg ml^{-1}), and Tween 80 (5 μg ml^{-1}) for 3 days followed by 48-h Gal induction and another 24-h incubation. Cells were harvested and refluxed in 20% KOH/50% EtOH for 10 min and extracted with petroleum ether three times. The organic phase was concentrated *in vacuo*. Gas chromatography–mass spectrometry (GC–MS) analysis was performed using an Agilent 7890A and Agilent 6540 Accurate-Mass Q-TOF (Santa Clara, USA). NMR analysis was performed on a Bruker AV 600 MHz spectrometer (Billerica, USA) (see supporting information Methods S1).

Selection analysis

The codon alignments and ML phylogenetic tree for putative OSCs from the previous step were used for selection analysis. We used HYPHY (v2.5.32) (http://hyphy.org/) to perform the BUSTED (Murrell et al., 2015), aBSREL (Smith et al., 2015), and MEME (Murrell et al., 2012) analyses. The 3D protein structure of *P. ginseng* CAS was downloaded from UniProt (UniProt Consortium, 2021) with identifier AF-O82139-F1 (predicted by AlphaFold [Jumper et al., 2021]). PyMOL was used for visualization of protein structures (The PyMOL Molecular Graphics System, Version 2.5, Schrödinger, LLC.).

ACCESSION NUMBERS

The raw sequencing data, assembly, and annotation files of *P. vietnamensis* var. *fuscidiscus* and *P. notoginseng* have been deposited at CNGBdb (https://db.cngb.org/) under project accession numbers CNP0002878 and CNP0003588.

SUPPLEMENTAL INFORMATION

Supplemental information is available at *Plant Communications Online*.

FUNDING

This work was supported by Digitalization of biological resources (202002AA100007), the Guangxi Innovation-Driven Development Project (GuiKe AA18242040), the General Project for Basic Research in Yunnan (grant no. 202201AT070266), and the National Natural Science Foundation of China (81860680).

AUTHOR CONTRIBUTIONS

S.Y. and Y.D. designed the research. Z.Y., X.L., L.Y., W.S., and G.Z. collected the data. Z.Y., X.L., L.Y., and S.P. performed the data and experimental analyses. Z.Y., W.C., S.Y., and Y.D. wrote the manuscript with contributions from all authors.

ACKNOWLEDGMENTS

We thank Pengchuan Sun from SiChuan University and Yue Wang for technical support. No conflict of interest is declared.

Received: August 23, 2022
Revised: January 14, 2023
Accepted: March 13, 2023
Published: March 16, 2023

REFERENCES

Aggarwal, G., and Ramaswamy, R. (2002). Ab initio gene identification: prokaryote genome annotation with GeneScan and GLIMMER. J. Biosci. **27**:7–14. https://doi.org/10.1007/BF02703679.

Alioto, T., Blanco, E., Parra, G., and Guigó, R. (2018). Using geneid to identify genes. Curr. Protoc. Bioinformatics **64**:e56. https://doi.org/10.1002/cpbi.56.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. **215**:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

Amborella Genome Project. (2013). The *Amborella* genome and the evolution of flowering plants. Science **342**:1241089. https://doi.org/10.1126/science.1241089.

Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. Mob. DNA **6**:11. https://doi.org/10.1186/s13100-015-0041-9.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. **27**:573–580. https://doi.org/10.1093/nar/27.2.573.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods **12**:59–60. https://doi.org/10.1038/nmeth.3176.

Busta, L., Schmitz, E., Kosma, D.K., Schnable, J.C., and Cahoon, E.B. (2021). A co-opted steroid synthesis gene, maintained in sorghum but not maize, is associated with a divergence in leaf wax chemistry. Proc. Natl. Acad. Sci. USA **118**, e2022982118. https://doi.org/10.1073/pnas.2022982118.

Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). EggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol. Biol. Evol. **38**:5825–5829. https://doi.org/10.1093/molbev/msab293.

Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics **25**:1972–1973. https://doi.org/10.1093/bioinformatics/btp348.

Cárdenas, P.D., Almeida, A., and Bak, S. (2019). Evolution of structural diversity of triterpenoids. Front. Plant Sci. **10**:1523. https://doi.org/10.3389/fpls.2019.01523.

Chen, C., Chen, H., Zhang, Y., Thomas, H.R., Frank, M.H., He, Y., and Xia, R. (2020). Tbtools: an integrative toolkit developed for interactive analyses of big biological data. Mol. Plant **13**:1194–1202. https://doi.org/10.1016/j.molp.2020.06.009.

Chen, K., Zhang, M., Ye, M., and Qiao, X. (2021a). Site-directed mutagenesis and substrate compatibility to reveal the

structure-function relationships of plant oxidosqualene cyclases. Nat. Prod. Rep. **38**:2261–2275. https://doi.org/10.1039/d1np00015b.

Chen, T., Zhang, H., Liu, Y., Liu, Y.X., and Huang, L. (2021b). EVenn: easy to create repeatable and editable Venn diagrams and Venn networks online. J. Genet. Genomics **48**:863–866. https://doi.org/10.1016/j.jgg.2021.07.007.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics **34**:i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

Delis, C., Krokida, A., Georgiou, S., Peña-Rodríguez, L.M., Kavroulakis, N., Ioannou, E., Roussis, V., Osbourn, A.E., and Papadopoulou, K.K. (2011). Role of lupeol synthase in *Lotus japonicus* nodule formation. New Phytol. **189**:335–346. https://doi.org/10.1111/j.1469-8137.2010.03463.x.

Dong, L., Almeida, A., Pollier, J., Khakimov, B., Bassard, J.E., Miettinen, K., Stærk, D., Mehran, R., Olsen, C.E., Motawia, M.S., et al. (2021). An independent evolutionary origin for insect deterrent cucurbitacins in *Iberis amara*. Mol. Biol. Evol. **38**:4659–4673. https://doi.org/10.1093/molbev/msab213.

Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., and Aiden, E.L. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science **356**:92–95. https://doi.org/10.1126/science.aal3327.

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. **3**:95–98. https://doi.org/10.1016/j.cels.2016.07.002.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics **14**:755–763. https://doi.org/10.1093/bioinformatics/14.9.755.

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics **9**:18. https://doi.org/10.1186/1471-2105-9-18.

Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. **20**:238. https://doi.org/10.1186/s13059-019-1832-y.

Fan, W., Huang, Y., Zheng, H., Li, S., Li, Z., Yuan, L., Cheng, X., He, C., and Sun, J. (2020). Ginsenosides for the treatment of metabolic syndrome and cardiovascular diseases: pharmacology and mechanisms. Biomed. Pharmacother. **132**:110915. https://doi.org/10.1016/j.biopha.2020.110915.

Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., and Smit, A.F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. Proc. Natl. Acad. Sci. USA **117**:9451–9457. https://doi.org/10.1073/pnas.1921046117.

Gas-Pascual, E., Berna, A., Bach, T.J., and Schaller, H. (2014). Plant oxidosqualene metabolism: cycloartenol synthase-dependent sterol biosynthesis in *Nicotiana benthamiana*. PLoS One **9**:e109156. https://doi.org/10.1371/journal.pone.0109156.

Gietz, R.D., and Schiestl, R.H. (2007). High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. Nat. Protoc. **2**:31–34. https://doi.org/10.1038/nprot.2007.13.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. **29**:644–652. https://doi.org/10.1038/nbt.1883.

Gremme, G., Brendel, V., Sparks, M.E., and Kurtz, S. (2005). Engineering a software tool for gene structure prediction in higher organisms. Inf. Softw. Technol. **47**:965–978. https://doi.org/10.1016/j.infsof.2005.09.005.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K., Jr., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., et al. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. **31**:5654–5666. https://doi.org/10.1093/nar/gkg770.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. **9**:R7. https://doi.org/10.1186/gb-2008-9-1-r7.

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. **35**:518–522. https://doi.org/10.1093/molbev/msx281.

Hou, M., Wang, R., Zhao, S., and Wang, Z. (2021). Ginsenosides in *Panax* genus and their biosynthesis. Acta Pharm. Sin. B **11**:1813–1834. https://doi.org/10.1016/j.apsb.2020.12.017.

Hu, J., Fan, J., Sun, Z., and Liu, S. (2020). NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics **36**:2253–2255. https://doi.org/10.1093/bioinformatics/btz891.

Huang, A.C., Jiang, T., Liu, Y.X., Bai, Y.C., Reed, J., Qu, B., Goossens, A., Nützmann, H.W., Bai, Y., and Osbourn, A. (2019). A specialized metabolic network selectively modulates *Arabidopsis* root microbiota. Science **364**:eaau6389. https://doi.org/10.1126/science.aau6389.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. **47**:D309–D314. https://doi.org/10.1093/nar/gky1085.

Inagaki, Y.S., Etherington, G., Geisler, K., Field, B., Dokarry, M., Ikeda, K., Mutsukado, Y., Dicks, J., and Osbourn, A. (2011). Investigation of the potential for triterpene synthesis in rice through genome mining and metabolic engineering. New Phytol. **191**:432–448. https://doi.org/10.1111/j.1469-8137.2011.03712.x.

Ito, R., Nakada, C., and Hoshino, T. (2016). β-Amyrin synthase from *Euphorbia tirucalli* L. functional analyses of the highly conserved aromatic residues Phe413, Tyr259 and Trp257 disclose the importance of the appropriate steric bulk, and cation-π and CH-π interactions for the efficient catalytic action of the polyolefin cyclization cascade. Org. Biomol. Chem. **15**:177–188. https://doi.org/10.1039/c6ob02539k.

Ji, Y., Liu, C., Yang, Z., Yang, L., He, Z., Wang, H., Yang, J., and Yi, T. (2019). Testing and using complete plastomes and ribosomal DNA sequences as the next generation DNA barcodes in *Panax* (Araliaceae). Mol. Ecol. Resour. **19**:1333–1345. https://doi.org/10.1111/1755-0998.13050.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature **596**:583–589. https://doi.org/10.1038/s41586-021-03819-2.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods **14**:587–589. https://doi.org/10.1038/nmeth.4285.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. **30**:772–780. https://doi.org/10.1093/molbev/mst010.

Khakimov, B., Kuzina, V., Erthmann, P.Ø., Fukushima, E.O., Augustin, J.M., Olsen, C.E., Scholtalbers, J., Volpin, H., Andersen, S.B., Hauser, T.P., et al. (2015). Identification and genome organization of

saponin pathway genes from a wild crucifer, and their use for transient production of saponins in *Nicotiana benthamiana*. Plant J. **84**:478–490. https://doi.org/10.1111/tpj.13012.

Kim, N.H., Jayakodi, M., Lee, S.C., Choi, B.S., Jang, W., Lee, J., Kim, H.H., Waminal, N.E., Lakshmanan, M., van Nguyen, B., et al. (2018a). Genome and evolution of the shade-requiring medicinal herb *Panax ginseng*. Plant Biotechnol. J. **16**:1904–1917. https://doi.org/10.1111/pbi.12926.

Kim, O.T., Um, Y., Jin, M.L., Kim, J.U., Hegebarth, D., Busta, L., Racovita, R.C., and Jetter, R. (2018b). A novel multifunctional C-23 Oxidase, CYP714E19, is involved in asiaticoside biosynthesis. Plant Cell Physiol. **59**:1200–1213. https://doi.org/10.1093/pcp/pcy055.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. **19**:1639–1645. https://doi.org/10.1101/gr.092759.109.

Kolesnikova, M.D., Xiong, Q., Lodeiro, S., Hua, L., and Matsuda, S.P.T. (2006). Lanosterol biosynthesis in plants. Arch. Biochem. Biophys. **447**:87–95. https://doi.org/10.1016/j.abb.2005.12.010.

Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. (2017). TimeTree: a resource for timelines, timetrees, and divergence times. Mol. Biol. Evol. **34**:1812–1819. https://doi.org/10.1093/molbev/msx116.

Kushiro, T., Shibuya, M., and Ebizuka, Y. (1999). Chimeric triterpene synthase. a possible model for multifunctional triterpene synthase. J. Am. Chem. Soc. **121**:1208–1216. https://doi.org/10.1021/ja983012h.

Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics **5**:59. https://doi.org/10.1186/1471-2105-5-59.

Landis, J.B., Soltis, D.E., Li, Z., Marx, H.E., Barker, M.S., Tank, D.C., and Soltis, P.S. (2018). Impact of whole-genome duplication events on diversification rates in angiosperms. Am. J. Bot. **105**:348–363. https://doi.org/10.1002/ajb2.1060.

Leung, K.W., and Wong, A.S.T. (2010). Pharmacology of ginsenosides: a literature review. Chin. Med. **5**:20. https://doi.org/10.1186/1749-8546-5-20.

Li, H., and Durbin, R. (2009a). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**:1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009b). The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Li, C., Xie, X., Li, F., Tian, E., Shu, Y., and Chao, Z. (2020). The complete chloroplast genome sequence of *Centella asiatica* (Linnaeus) Urban. Mitochondrial DNA. B Resour. **5**:2149–2150. https://doi.org/10.1080/23802359.2020.1768922.

Li, M.R., Ding, N., Lu, T., Zhao, J., Wang, Z.H., Jiang, P., Liu, S.T., Wang, X.F., Liu, B., and Li, L.F. (2021a). Evolutionary contribution of duplicated genes to genome evolution in the ginseng species complex. Genome Biol. Evol. **13**:evab051. https://doi.org/10.1093/gbe/evab051.

Li, Y., Leveau, A., Zhao, Q., Feng, Q., Lu, H., Miao, J., Xue, Z., Martin, A.C., Wegel, E., Wang, J., et al. (2021b). Subtelomeric assembly of a multi-gene pathway for antimicrobial defense compounds in cereals. Nat. Commun. **12**:2563. https://doi.org/10.1038/s41467-021-22920-8.

Lichman, B.R., Godden, G.T., and Buell, C.R. (2020). Gene and genome duplications in the evolution of chemodiversity: perspectives from studies of Lamiaceae. Curr. Opin. Plant Biol. **55**:74–83. https://doi.org/10.1016/j.pbi.2020.03.005.

Ma, Y., Zhou, Y., Ovchinnikov, S., Greisen, P., Jr., Huang, S., and Shang, Y. (2016). New insights into substrate folding preference of plant OSCs. Sci. Bull. **61**:1407–1412. https://doi.org/10.1007/s11434-016-1103-1.

Majoros, W.H., Pertea, M., and Salzberg, S.L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics **20**:2878–2879. https://doi.org/10.1093/bioinformatics/bth315.

Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., and Zdobnov, E.M. (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol. Biol. Evol. **38**:4647–4654. https://doi.org/10.1093/molbev/msab199.

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics **27**:764–770. https://doi.org/10.1093/bioinformatics/btr011.

Mendes, F.K., Vanderpool, D., Fulton, B., and Hahn, M.W. (2020). CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics **36**:5516–5518. https://doi.org/10.1093/bioinformatics/btaa1022.

Miettinen, K., Iñigo, S., Kreft, L., Pollier, J., De Bo, C., Botzki, A., Coppens, F., Bak, S., and Goossens, A. (2018). The TriForC database: a comprehensive up-to-date resource of plant triterpene biosynthesis. Nucleic Acids Res. **46**:D586–D594. https://doi.org/10.1093/nar/gkx925.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: the protein families database in 2021. Nucleic Acids Res. **49**:D412–D419. https://doi.org/10.1093/nar/gkaa913.

Morita, M., Shibuya, M., Lee, M.S., Sankawa, U., and Ebizuka, Y. (1997). Molecular cloning of pea cDNA encoding cycloartenol synthase and its functional expression in yeast. Biol. Pharm. Bull. **20**:770–775. https://doi.org/10.1248/bpb.20.770.

Murrell, B., Wertheim, J.O., Moola, S., Weighill, T., Scheffler, K., and Kosakovsky Pond, S.L. (2012). Detecting individual sites subject to episodic diversifying selection. PLoS Genet. **8**:e1002764. https://doi.org/10.1371/journal.pgen.1002764.

Murrell, B., Weaver, S., Smith, M.D., Wertheim, J.O., Murrell, S., Aylward, A., Eren, K., Pollner, T., Martin, D.P., Smith, D.M., et al. (2015). Gene-wide identification of episodic selection. Mol. Biol. Evol. **32**:1365–1371. https://doi.org/10.1093/molbev/msv035.

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**:418–426. https://doi.org/10.1093/oxfordjournals.molbev.a040410.

Neumann, P., Novák, P., Hoštáková, N., and Macas, J. (2019). Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. Mob. DNA **10**:1. https://doi.org/10.1186/s13100-018-0144-1.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. **32**:268–274. https://doi.org/10.1093/molbev/msu300.

Noda-Garcia, L., and Tawfik, D.S. (2020). Enzyme evolution in natural products biosynthesis: target- or diversity-oriented? Curr. Opin. Chem. Biol. **59**:147–154. https://doi.org/10.1016/j.cbpa.2020.05.011.

Ober, D. (2005). Seeing double: gene duplication and diversification in plant secondary metabolism. Trends Plant Sci. **10**:444–449. https://doi.org/10.1016/j.tplants.2005.07.007.

Ou, S., Chen, J., and Jiang, N. (2018a). Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Res. **46**:e126. https://doi.org/10.1093/nar/gky730.

Ou, S., and Jiang, N. (2018b). LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. **176**:1410–1422. https://doi.org/10.1104/pp.17.01310.

Ou, S., and Jiang, N. (2019). LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. Mob. DNA **10**:48. https://doi.org/10.1186/s13100-019-0193-0.

Pichersky, E., and Lewinsohn, E. (2011). Convergent evolution in plant specialized metabolism. Annu. Rev. Plant Biol. **62**:549–566. https://doi.org/10.1146/annurev-arplant-042110-103814.

Qin, L., Hu, Y., Wang, J., Wang, X., Zhao, R., Shan, H., Li, K., Xu, P., Wu, H., Yan, X., et al. (2021). Insights into angiosperm evolution, floral development and chemical biosynthesis from the *Aristolochia fimbriata* genome. Nat. Plants **7**:1239–1253. https://doi.org/10.1038/s41477-021-00990-2.

Ranallo-Benavidez, T.R., Jaron, K.S., and Schatz, M.C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat. Commun. **11**:1432. https://doi.org/10.1038/s41467-020-14998-3.

Roach, M.J., Schmidt, S.A., and Borneman, A.R. (2018). Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinf. **19**:460. https://doi.org/10.1186/s12859-018-2485-7.

Santana-Molina, C., Rivas-Marin, E., Rojas, A.M., and Devos, D.P. (2020). Origin and evolution of polycyclic triterpene synthesis. Mol. Biol. Evol. **37**:1925–1941. https://doi.org/10.1093/molbev/msaa054.

Schaller, H. (2003). The role of sterols in plant growth and development. Prog. Lipid Res. **42**:163–175. https://doi.org/10.1016/s0163-7827(02)00047-4.

Smith, M.D., Wertheim, J.O., Weaver, S., Murrell, B., Scheffler, K., and Kosakovsky Pond, S.L. (2015). Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Mol. Biol. Evol. **32**:1342–1353. https://doi.org/10.1093/molbev/msv022.

Song, X., Sun, P., Yuan, J., Gong, K., Li, N., Meng, F., Zhang, Z., Li, X., Hu, J., Wang, J., et al. (2021). The celery genome sequence reveals sequential paleo-polyploidizations, karyotype evolution and resistance gene reduction in Apiales. Plant Biotechnol. J. **19**:731–744. https://doi.org/10.1111/pbi.13499.

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics **24**:637–644. https://doi.org/10.1093/bioinformatics/btn013.

Stull, G.W., Qu, X.J., Parins-Fukuchi, C., Yang, Y.Y., Yang, J.B., Yang, Z.Y., Hu, Y., Ma, H., Soltis, P.S., Soltis, D.E., et al. (2021). Gene duplications and phylogenomic conflict underlie major pulses of phenotypic evolution in gymnosperms. Nat. Plants **7**:1015–1025. https://doi.org/10.1038/s41477-021-00964-4.

Su, W., Jing, Y., Lin, S., Yue, Z., Yang, X., Xu, J., Wu, J., Zhang, Z., Xia, R., Zhu, J., et al. (2021). Polyploidy underlies co-option and diversification of biosynthetic triterpene pathways in the apple tribe. Proc. Natl. Acad. Sci. USA **118**, e2101767118. https://doi.org/10.1073/pnas.2101767118.

1. **Sun, P., Jiao, B., Yang, Y., et al.** (2022). WGDI: a user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotype. Mol. Plant. **15**:1841–1851. https://doi.org/10.1016/j.molp.2022.10.018.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. **34**:W609–W612. https://doi.org/10.1093/nar/gkl315.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., and Paterson, A.H. (2008). Synteny and collinearity in plant genomes. Science **320**:486–488. https://doi.org/10.1126/science.1153917.

Tansakul, P., Shibuya, M., Kushiro, T., and Ebizuka, Y. (2006). Dammarenediol-II synthase, the first dedicated enzyme for ginsenoside biosynthesis, in Panax ginseng. FEBS Lett. **580**:5143–5149. https://doi.org/10.1016/j.febslet.2006.08.044.

Tawfik, D.S., and Gruic-Sovulj, I. (2020). How evolution shapes enzyme selectivity - lessons from aminoacyl-tRNA synthetases and other amino acid utilizing enzymes. FEBS J. **287**:1284–1305. https://doi.org/10.1111/febs.15199.

Thimmappa, R., Geisler, K., Louveau, T., O'Maille, P., and Osbourn, A. (2014). Triterpene biosynthesis in plants. Annu. Rev. Plant Biol. **65**:225–257. https://doi.org/10.1146/annurev-arplant-050312-120229.

Tien, N.Q.D., Ma, X., Man, L.Q., et al. (2021). De novo whole-genome assembly and discovery of genes involved in triterpenoid saponin biosynthesis of Vietnamese ginseng (*Panax vietnamensis* Ha *et* Grushv.). Physiol. Mol. Biol. Plants **27**:2215–2229. https://doi.org/10.1007/s12298-021-01076-1.

Tokuriki, N., Jackson, C.J., Afriat-Jurnou, L., Wyganowski, K.T., Tang, R., and Tawfik, D.S. (2012). Diminishing returns and tradeoffs constrain the laboratory optimization of an enzyme. Nat. Commun. **3**:1257. https://doi.org/10.1038/ncomms2246.

UniProt Consortium. (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. **49**:D480–D489. https://doi.org/10.1093/nar/gkaa1100.

van Dongen, S. (2000). Graph clustering by flow simulation (University of Utrecht). PhD thesis.

Wang, J., Guo, Y., Yin, X., Wang, X., Qi, X., and Xue, Z. (2022). Diverse triterpene skeletons are derived from the expansion and divergent evolution of 2,3-oxidosqualene cyclases in plants. Crit. Rev. Biochem. Mol. Biol. **57**:113–132. https://doi.org/10.1080/10409238.2021.1979458.

Wang, L., Zhao, S.J., Cao, H.J., et al. (2014). The isolation and characterization of dammarenediol synthase gene from Panax quinquefolius and its heterologous co-expression with cytochrome P450 gene PqD12H in yeast. Funct. Integr. Genomics **14**:545–557. https://doi.org/10.1007/s10142-014-0384-1.

Wu, T., Liu, W., Huang, S., Chen, J., He, F., Wang, H., Zheng, X., Li, Z., Zhang, H., Zha, Z., et al. (2021). clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Mater. Today. Bio **12**:100141. https://doi.org/10.1080/10409238.2021.1979458.

Xue, Z., Duan, L., Liu, D., Guo, J., Ge, S., Dicks, J., ÓMáille, P., Osbourn, A., and Qi, X. (2012). Divergent evolution of oxidosqualene cyclases in plants. New Phytol. **193**:1022–1038. https://doi.org/10.1111/j.1469-8137.2011.03997.x.

Yang, Z. (2007). Paml 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24**:1586–1591. https://doi.org/10.1093/molbev/msm088.

Yang, Z., Liu, G., Zhang, G., Yan, J., Dong, Y., Lu, Y., Fan, W., Hao, B., Lin, Y., Li, Y., et al. (2021a). The chromosome-scale high-quality genome assembly of *Panax notoginseng* provides insight into dencichine biosynthesis. Plant Biotechnol. J. **19**:869–871. https://doi.org/10.1111/pbi.13558.

Yang, Z., Chen, S., Wang, S., Hu, Y., Zhang, G., Dong, Y., Yang, S., Miao, J., Chen, W., and Sheng, J. (2021b). Chromosomal-scale genome assembly of *Eleutherococcus senticosus* provides insights into chromosome evolution in Araliaceae. Mol. Ecol. Resour. **21**:2204–2220. https://doi.org/10.1111/1755-0998.13403.

Yu, G., Lam, T.T.Y., Zhu, H., and Guan, Y. (2018). Two methods for mapping and visualizing associated data on phylogeny using ggtree. Mol. Biol. Evol. **35**:3041–3043. https://doi.org/10.1093/molbev/msy194.

Zhang, G.H., Ma, C.H., Zhang, J.J., Chen, J.W., Tang, Q.Y., He, M.H., Xu, X.Z., Jiang, N.H., and Yang, S.C. (2015). Transcriptome analysis of *Panax vietnamensis* var. *fuscidiscus* discovers putative ocotillol-type ginsenosides biosynthesis genes and genetic markers. BMC Genom. **16**:159. https://doi.org/10.1186/s12864-015-1332-8.

Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics **19**:153. https://doi.org/10.1186/s12859-018-2129-y.

Zhang, R.G., Li, G.Y., Wang, X.L., Dainat, J., Wang, Z.X., Ou, S., and Ma, Y. (2022). TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. Hortic. Res. **9**:uhac017. https://doi.org/10.1093/hr/uhac017.

Zhou, Y., Ma, Y., Zeng, J., Duan, L., Xue, X., Wang, H., Lin, T., Liu, Z., Zeng, K., Zhong, Y., et al. (2016). Convergence and divergence of bitterness biosynthesis and regulation in Cucurbitaceae. Nat. Plants **2**:16183. https://doi.org/10.1038/nplants.2016.183.

Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L., and Gruissem, W. (2004). GENEVESTIGATOR. *Arabidopsis* microarray database and analysis toolbox. Plant Physiol. **136**:2621–2632. https://doi.org/10.1104/pp.104.046367.

Plant Communications, Volume *4*

Supplemental information

Comparative genomics reveals the diversification of triterpenoid biosynthesis and origin of ocotillol-type triterpenes in *Panax*

Zijiang Yang, Xiaobo Li, Ling Yang, Sufang Peng, Wanling Song, Yuan Lin, Guisheng Xiang, Ying Li, Shuang Ye, Chunhua Ma, Jianhua Miao, Guanghui Zhang, Wei Chen, Shengchao Yang, and Yang Dong

Zijiang Yang, Xiaobo Li, Ling Yang, Wanling Song, Yuan Lin, Guisheng Xiang, Ying Li, Shuang Ye, Chunhua Ma, Jianhua Miao, Guanghui Zhang, Wei Chen, Shengchao Yang, Yang Dong.

National & Local Joint Engineering Research Center on Germplasm Innovation & Utilization of Chinese Medicinal Materials in Southwest China, Yunnan Agricultural University, Kunming, China

The Key Laboratory of Medicinal Plant Biology of Yunnan Province, Yunnan Agricultural University, Kunming, China

College of Food Science and Technology, Yunnan Agricultural University, Kunming, China

Guangxi Key Laboratory of Medicinal Resources Protection and Genetic Improvement, Guangxi Botanical Garden of Medicinal Plants, Nanning, China

Yunnan Plateau Characteristic Agriculture Industry Research Institute, Kunming, China

Comparative genomics reveals the diversification of triterpenoid biosynthesis and origin of ocotillol-type triterpenes in *Panax*

Supporting Information Methods S1. Characterization of compounds produced from enzymatic reactions

Gas chromatography-mass spectrometry (GC-MS) analysis

The purified yeast extract was derivatized by resuspending in 100 μl of trimethylsilyl cyanide (TMSCN) with a 1:1 ratio followed by incubation of 30 min at 65 ℃. GC analysis was performed by Agilent 7890A with a HP-5MS quartz capillary column (30 m×0.25 mm i.d., 0.25 μm film thickness). The temperature was set as 250 ℃ for injector port, source, and transfer line. The column temperature was programmed as follows: 80 ℃ for 2 min; increase to 290 ℃ at a rate of 20 ℃ min^{-1}; hold at 290 ℃ for 30 min. The flow rate of carrier gas helium was 1.2 ml min^{-1}. Samples were injected in splitless mode with either a 1-μl or a 3-μl volume. MS analysis was performed using Agilent 6540 Accurate-Mass Q-TOF system.

Triterpenoid standards preparation

δ-amyrin, β-amyrin, α-amyrin, cycloartenol, ψ-taraxasterol, taraxasterol, and dammarendiol-II were purchased from Chengdu DeSiTe Biological Technology Co. Ltd, China; 3-epicabraleadiol was purchased from BioBioPha Co. Ltd, China. Standards were first dissolved in hexane, followed by derivatization using TMSCN.

Nuclear Magnetic Resonance (NMR) analysis

The purified yeast extract was subjected to column chromatography (CC) on silica gel (200-300 mesh, Qingdao Marine Chemical Factory, China) eluting with petroleum ether and then with petroleum ether/ethyl acetate stepwise-gradient system (from 13:1 to 5:1, v/v) to obtain four fractions (denoted as Fr.1–Fr.4). Fr.4 was purified by semi-preparative high-performance liquid chromatography (HPLC) on Agilent 1290 Infinity II system with off-line monitoring by thin-layer chromatography (TLC). The column used for HPLC was a reversed-phase column (Agilent ZORBAX StableBond SB-C18, 9.4 × 250 mm, 5 μm). The setting for mobile phase was 100% acetonitrile at a flow rate of 3 ml min^{-1}. TLC analysis was carried out on silica gel plates (GF254F, 10 - 40 μm, Qingdao Marine Chemical Factory) by spraying with 5% H2SO4 in EtOH (v/v) followed by heating to 120 °C for 5 min. The above process yielded compound **8** (3 mg, containing trace amount of compound **9**) and compound **9** (12 mg). The purified compound 8 and 9 was analyzed by ^{1}H-NMR and ^{13}C-NMR spectroscopy at 600 and 150 MHz in $CDCl_3$ solution using Bruker AV-600 MHz spectrometer.

Identification of compounds

Through GC analysis, the product profile for nine OSCs were identified (Table S24). The naming of compounds was in consistent with Figure 4C. Based on the GC retention times and mass spectral fragmentation patterns from existing literatures (Shan *et al.*, 2005; Salmon *et al.*, 2016; Kim *et al.*, 2018), the compounds were identified as follows: compound **1**: δ-amyrin; compound **2**: β-amyrin; compound **3**: α-amyrin; compound **5**: ψ-taraxasterol, compound **6**: taraxasterol; compound **7**: dammarendiol-II; compound **8**: 3-epicabraleadiol; compound **9**: ocotillol (Figures S19-S22).

Since authentic standard for ocotillol is not available. The NMR analysis was further performed to characterize the compound **8** and compound **9**. By comparison of NMR and mass spectroscopic data with previous study (Shan *et al.*, 2005), C-24S or C-24R epimers of the epoxydammaranes

can be distinguished by the ^{1}H-NMR chemical shifts of H-24 and ^{13}C-NMR chemical shifts of C-24, C-25, C-21, C-23 and C-22 positions. Chemical shifts and coupling constants at H-24 vary remarkably for molecules with locally diastereomeric configurations at C-20 and C-24 (Figures S24-S26, Table S23).

References for Methods S1

Shan, H., Segura, M.J., Wilson, W.K., Lodeiro, S., Matsuda, S.P. (2005). Enzymatic cyclization of dioxidosqualene to heterocyclic triterpenes. J. Am. Chem. Soc. **127**: 18008–18009.

Salmon, M., Thimmappa, R.B., Minto, R.E., Melton, R.E., Hughes, R.K., O'Maille, P.E., Hemmings, A.M., Osbourn, A. (2016). A conserved amino acid residue critical for product and substrate specificity in plant triterpene synthases. Proc. Natl. Acad. Sci. U.S.A. **113**: E4407–E4414.

Kim, O.T., Um, Y., Jin, M.L., Kim, J.U., Hegebarth, D., Busta, L., Racovita, R.C., Jetter, R. (2018). A Novel Multifunctional C-23 Oxidase, CYP714E19, is Involved in Asiaticoside Biosynthesis. Plant Cell Physiol. **59**: 1200–1213.

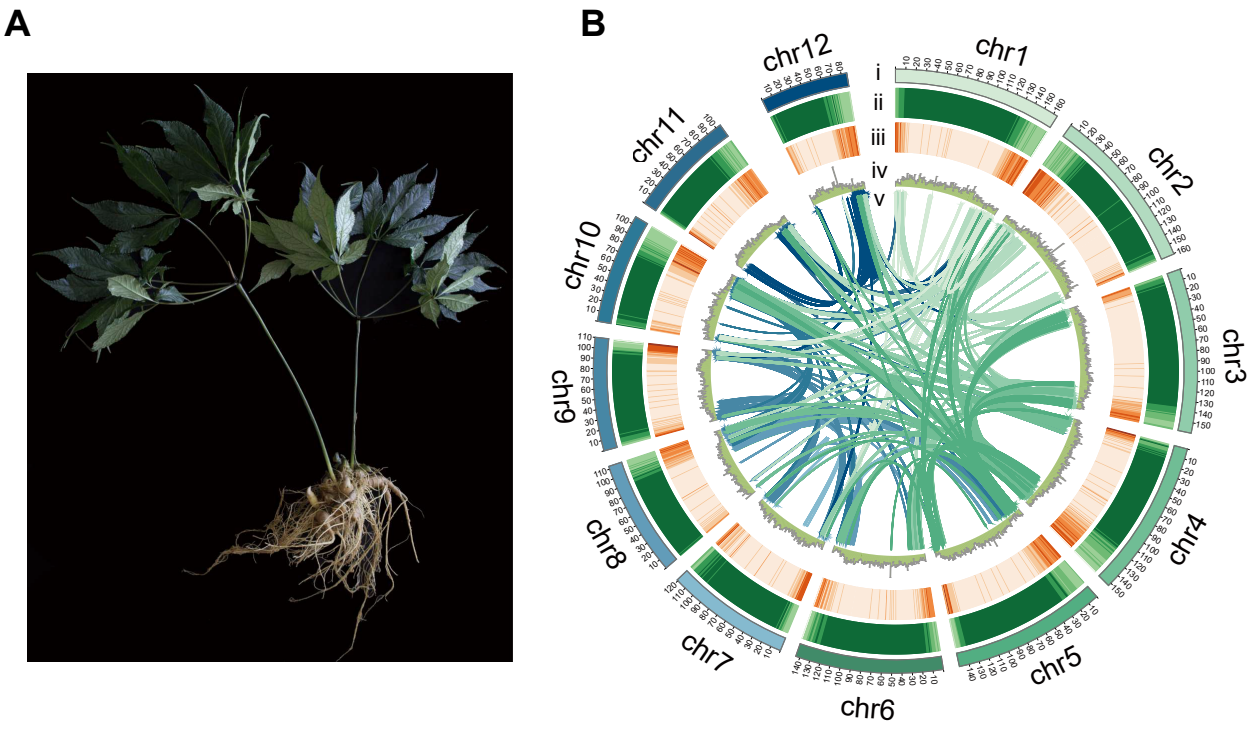

Figure S1 Morphology and genome of *P. vietnamensis* var. *fuscidiscus*. (A) Morphology of *P. vietnamensis* var. *fuscidiscus*. (B) Overview of *P. vietnamensis* var. *fuscidiscus* assembly. (I) chromosomes; (II) transposable elements density heatmap (1 Mb sliding window); (III) gene density heatmap (1 Mb sliding window); (IV) GC content (1 Mb sliding window); (V) collinear regions within *P. vietnamensis* var. *fuscidiscus* genome.

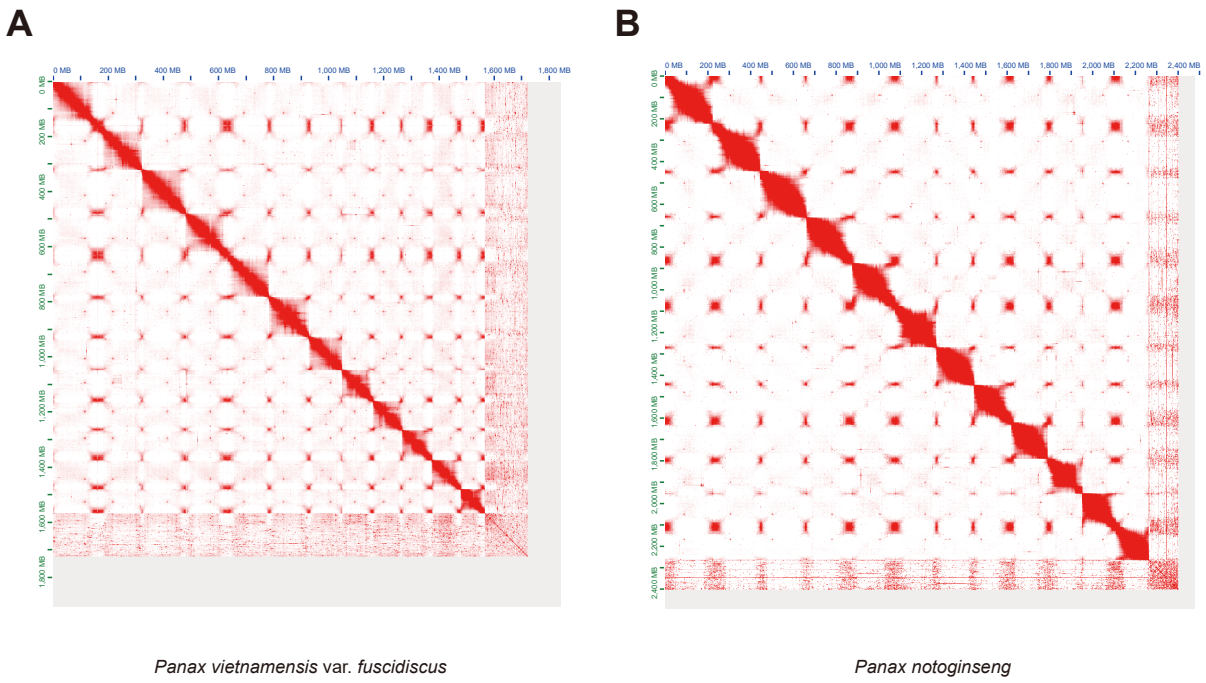

Figure S2 Hi-C contact heatmaps. (A) *P. vietnamensis* var. *fuscidiscus* assembly. (B) Updated *P. notoginseng* assembly.

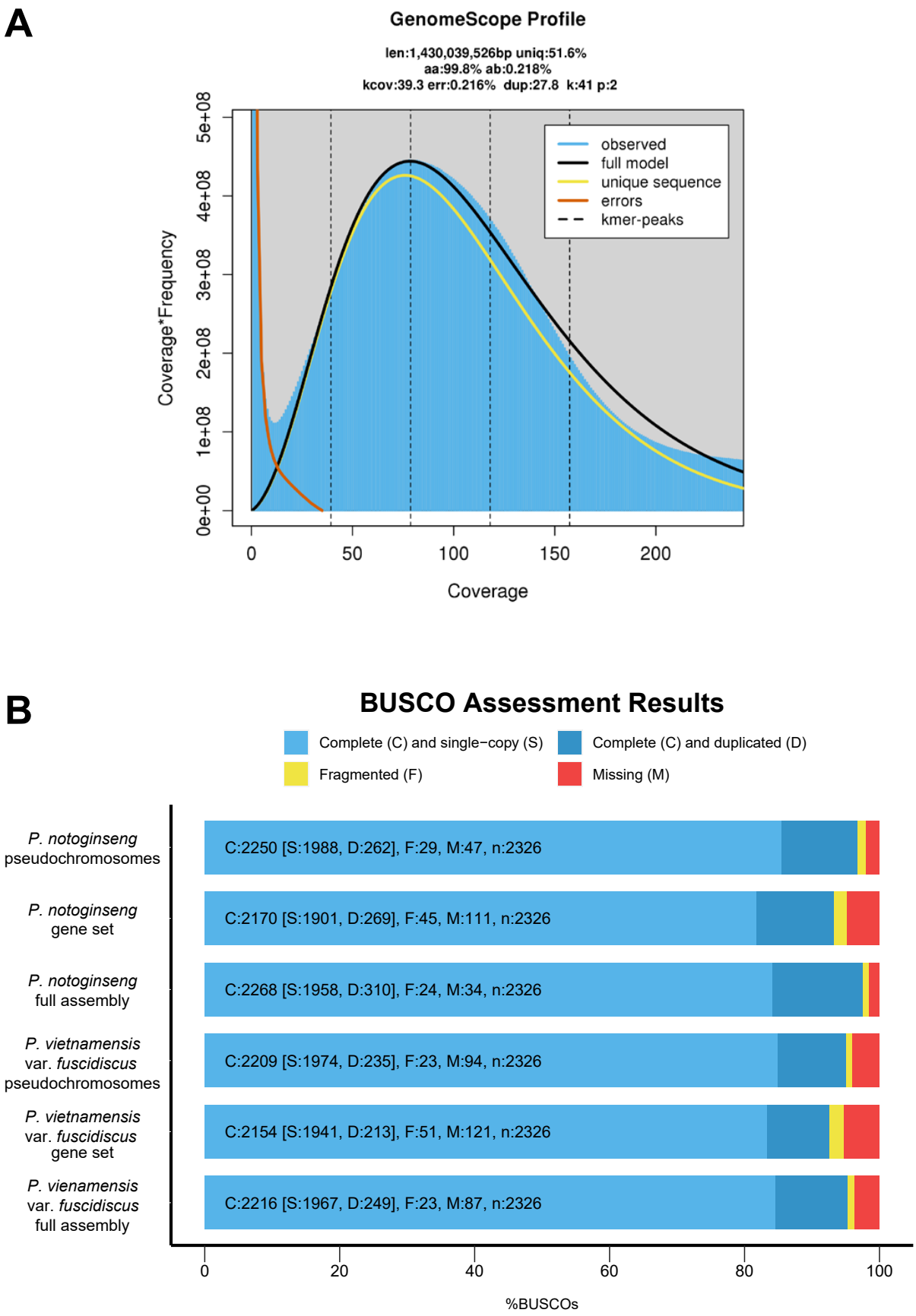

Figure S3 Genome survey and evaluation. (A) Survey of *P. vietnamensis* var. *fuscidiscus* genome by Genomescope. Ploidy and kmer length were set as 2 and 41, respectively. (B) BUSCO analysis of *P. vietnamensis* var. *fuscidiscus* and *P. notoginseng* assemblies and gene sets using the lineage dataset eudicots_odb10 (2,326 BUSCOs).

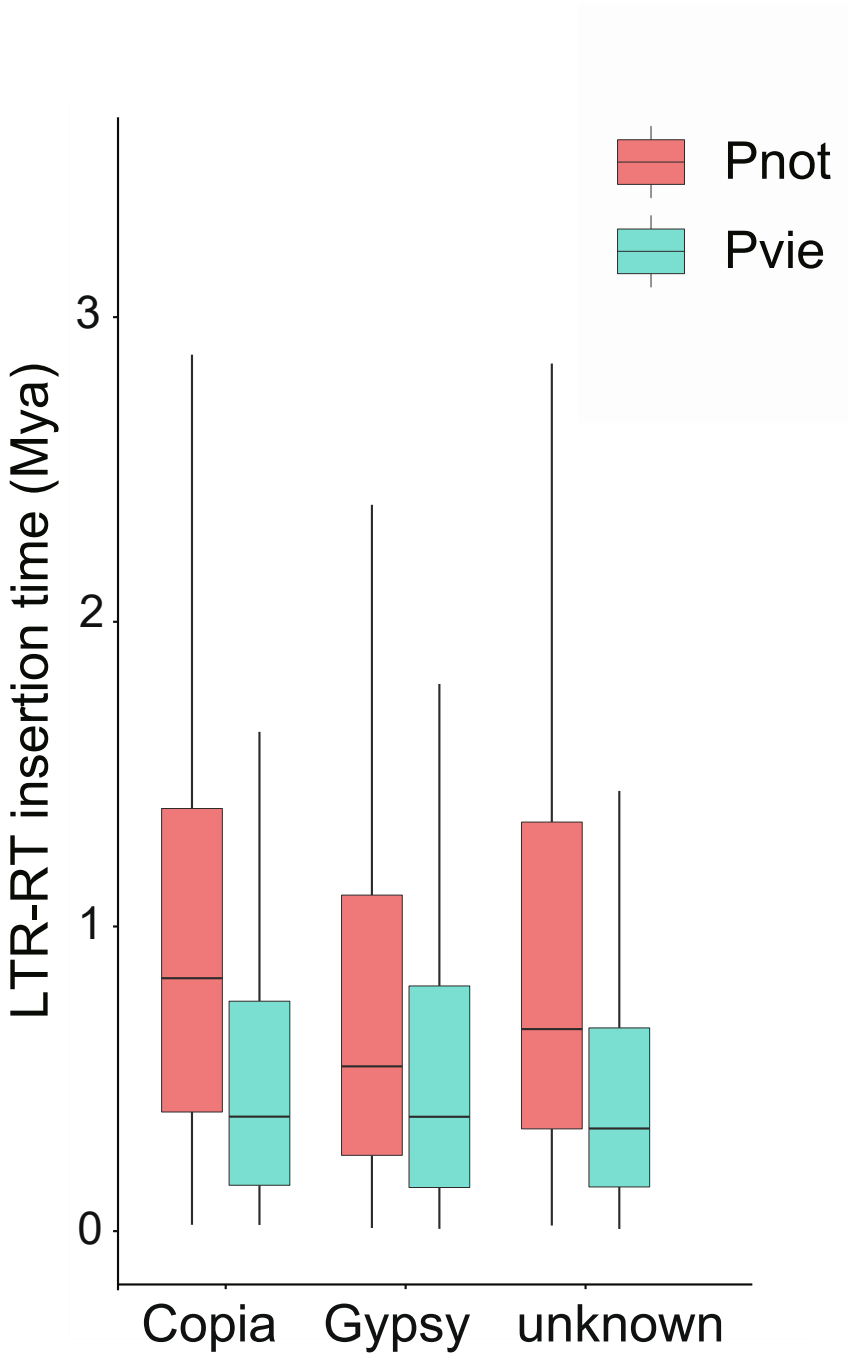

Figure S4 Estimated insertion time for Copia, Gypsy, and unknown type of LTRs in *P. vietnamensis* var. *fuscidiscus* and *P. notoginseng* genomes. Pvie: *P. vietnamensis* var. *fuscidiscus*, Pnot: *P. notoginseng*.

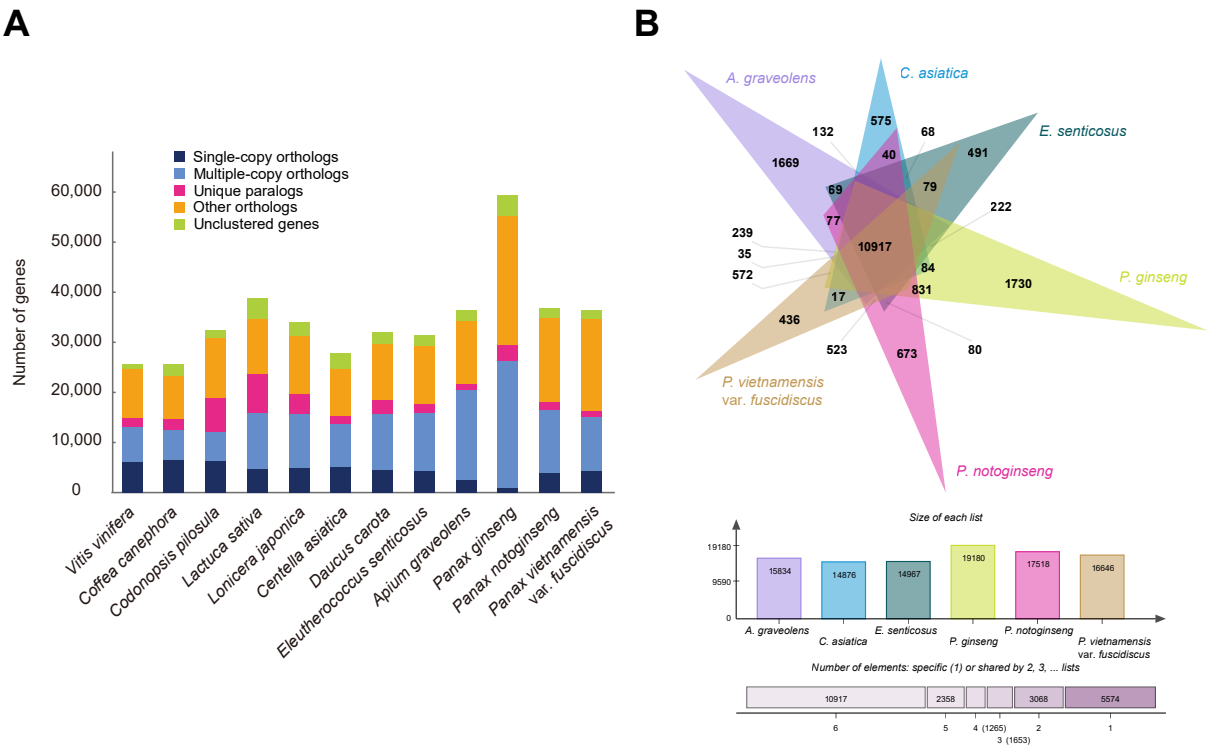

Figure S5 Orthology analyses. (A) Gene family analysis result for 12 eudicots. (B) Comparison of orthogroups from six Apiales species. Numbers in the upper venn diagram indicate the number of orthogroups. Numbers in the middle bar plot indicate the genes in the orthogroups for six species. Numbers in the bottom chart indicate number of shared or species-specific orthogroups.

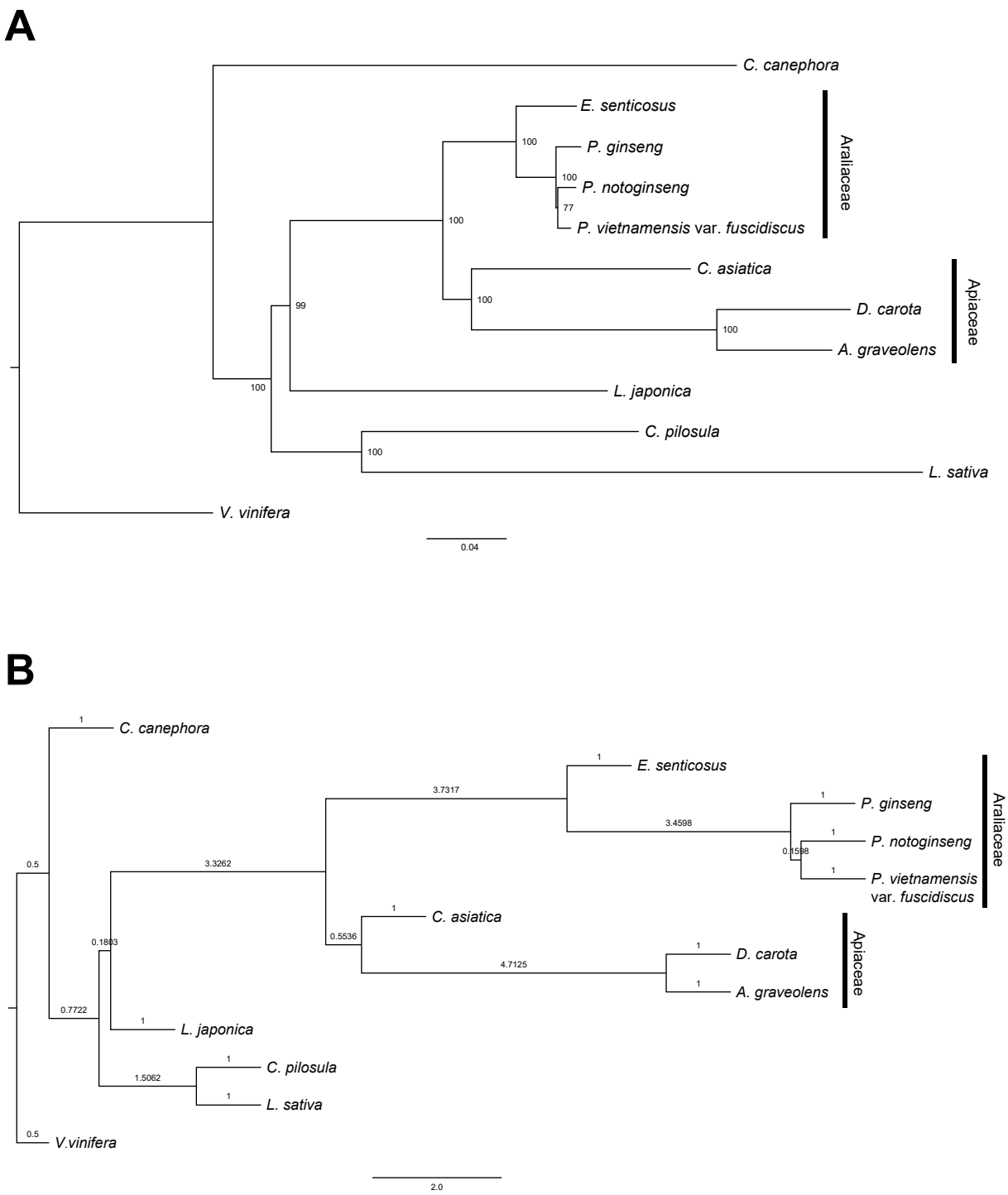

Figure S6 Phylogenetic analyses. (A) Species tree for 12 species inferred by concatenation-based method. Numbers indicate bootstrap values with 1,000 replicates. (B) Species tree for 12 species inferred by coalescence-based method. Branch lengths are shown in coalescent units. The numbers of each node represents the local posterior probabilities. Since the ASTRAL tree leaves the branch length of terminal branches empty, the length of terminal branches were all set as one.

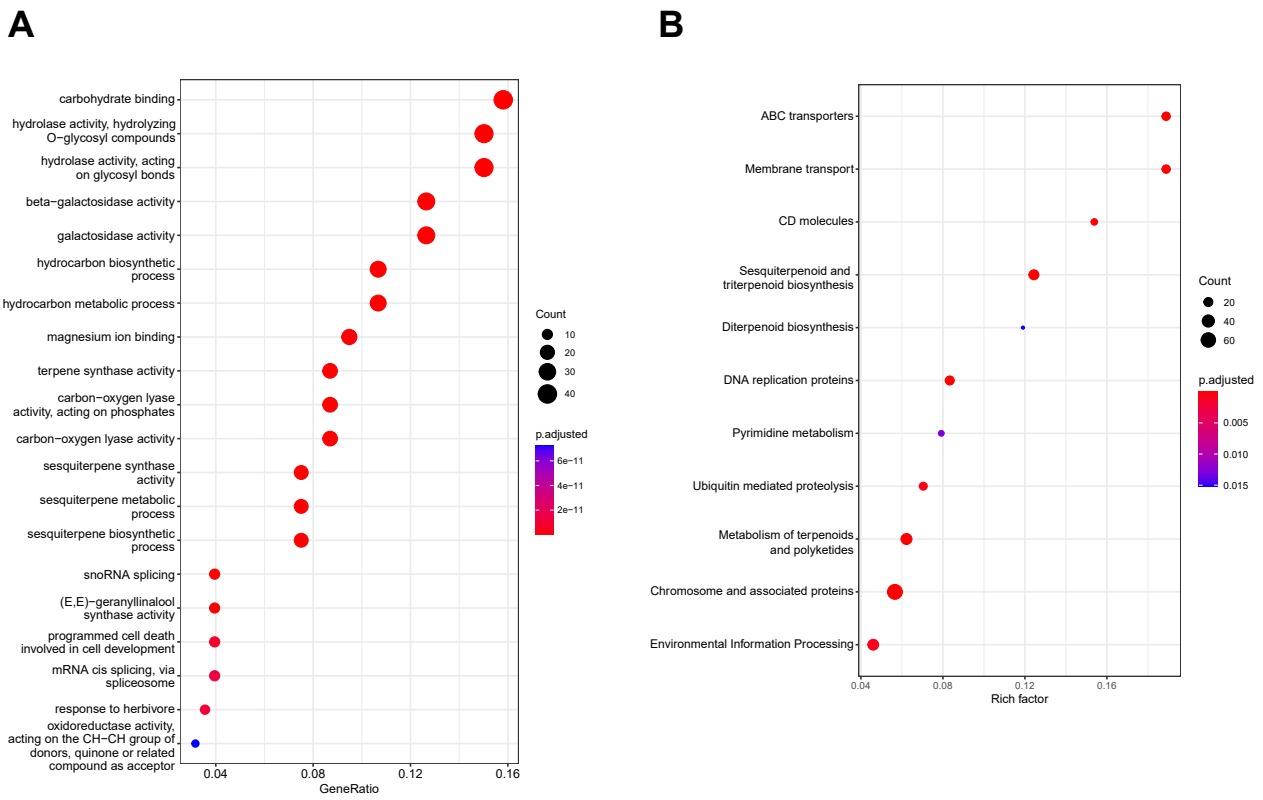

Figure S7 Functional enrichment analyses. (A) GO enrichment analysis of expanded gene families in *P. vietnamensis* var. *fuscidiscus*. (B) KEGG enrichment analysis of expanded gene families in *P. vietnamensis* var. *fuscidiscus*.

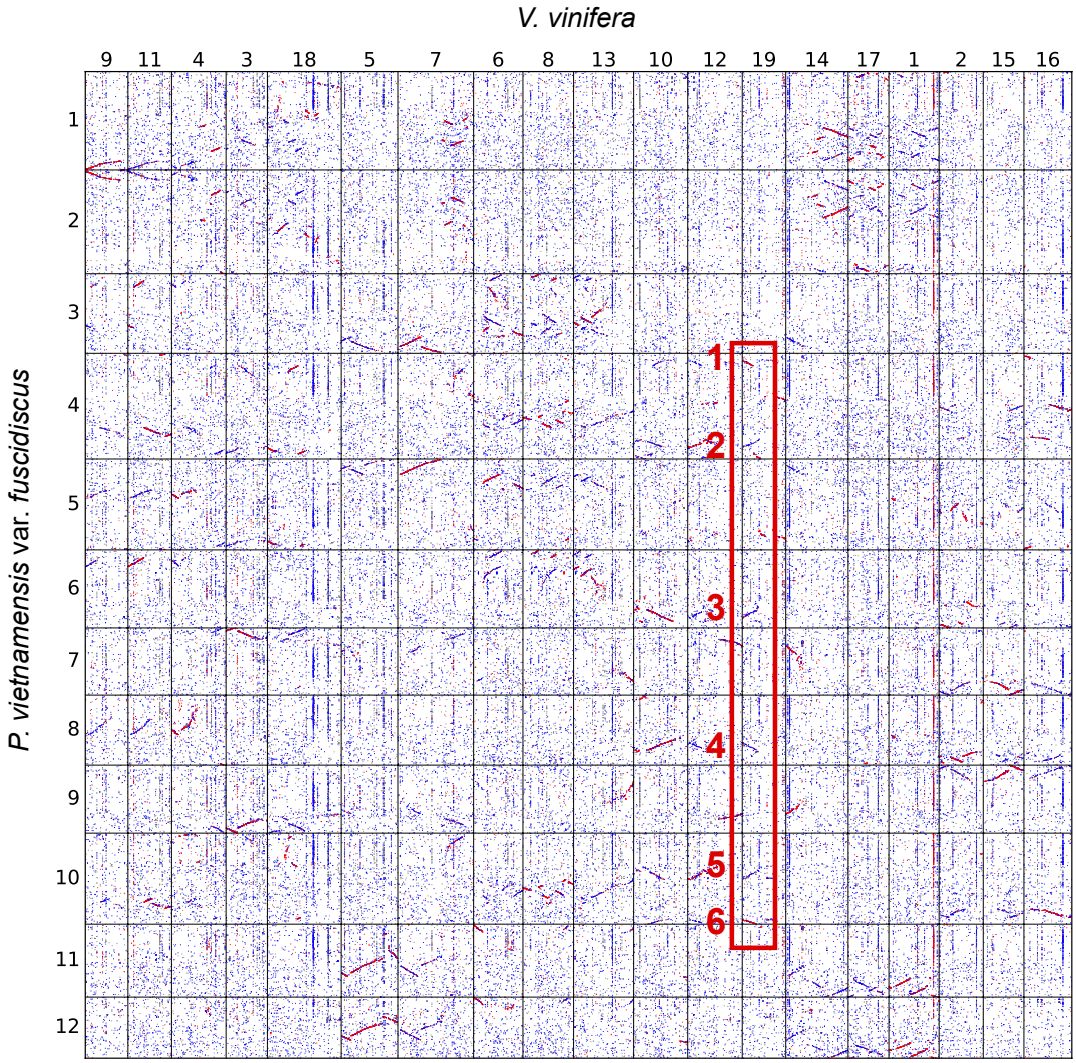

Figure S8 Synteny dot plot between *P. vietnamensis* var. *fuscidiscus* and *V. vinifera*. The red box highlighted regions between *V. vinifera* and *P. vietnamensis* var. *fuscidiscus* with a ratio of 1:6.

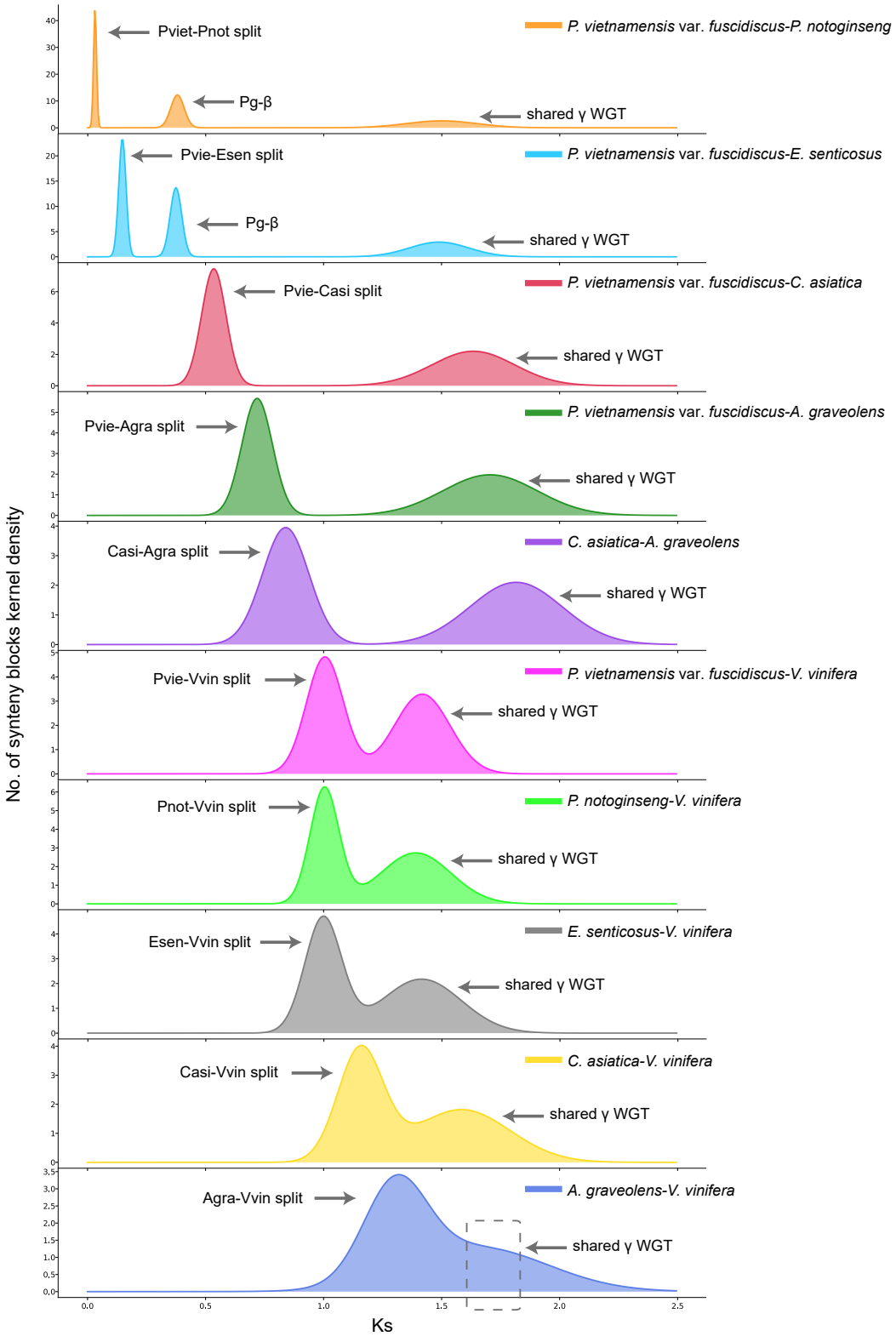

Figure S9 *Ks* distribution of interspecific collinear blocks. *Ks* peaks of speciation and shared polyploidizations between the studies species were labeled. Pvie: *P. vietnamensis* var. *fuscidiscus*, Pnot: *P. notoginseng*, Esen: *E. senticosus*, Casi: *C. asiatica*, Agra: *A. graveolens*, Vvin: *V. vinifera*.

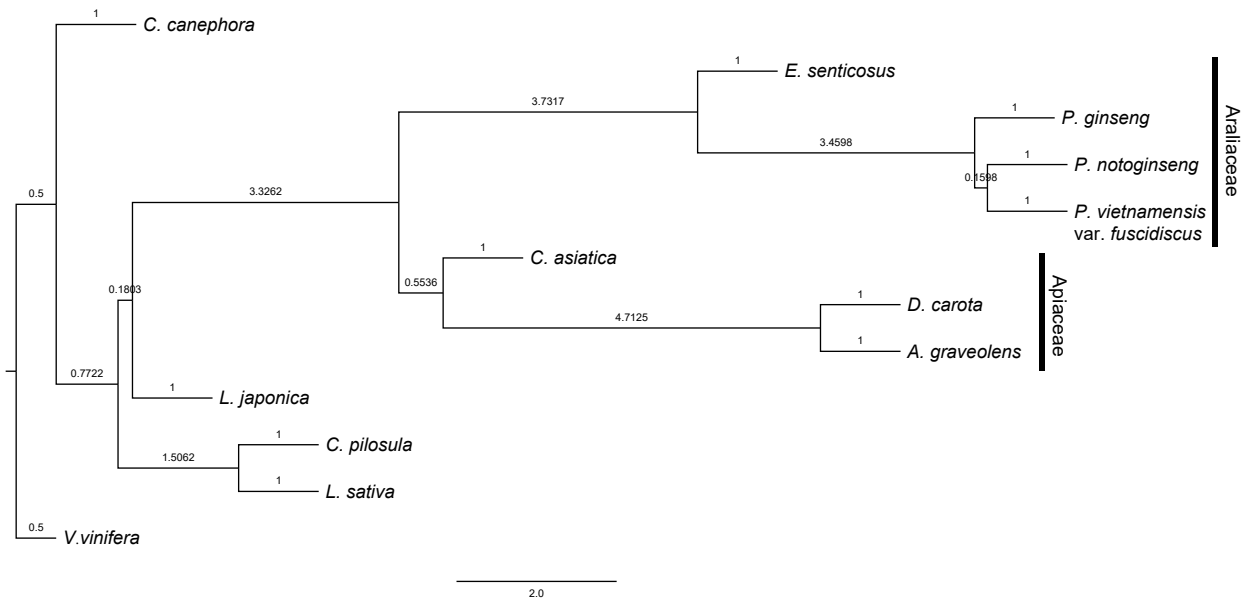

Figure S10 Synteny analyses. (A) Synteny dot plot between *P. notoginseng* and *E. senticosus*. *P. notoginseng* genome version PBJ-2021 (Yang *et al.*, 2021a). For *P. notoginseng*, number 13-18 represent Scaffold 13-18. (B) Synteny dot plot between *P. notoginseng* (updated by this study) and *E. senticosus*.

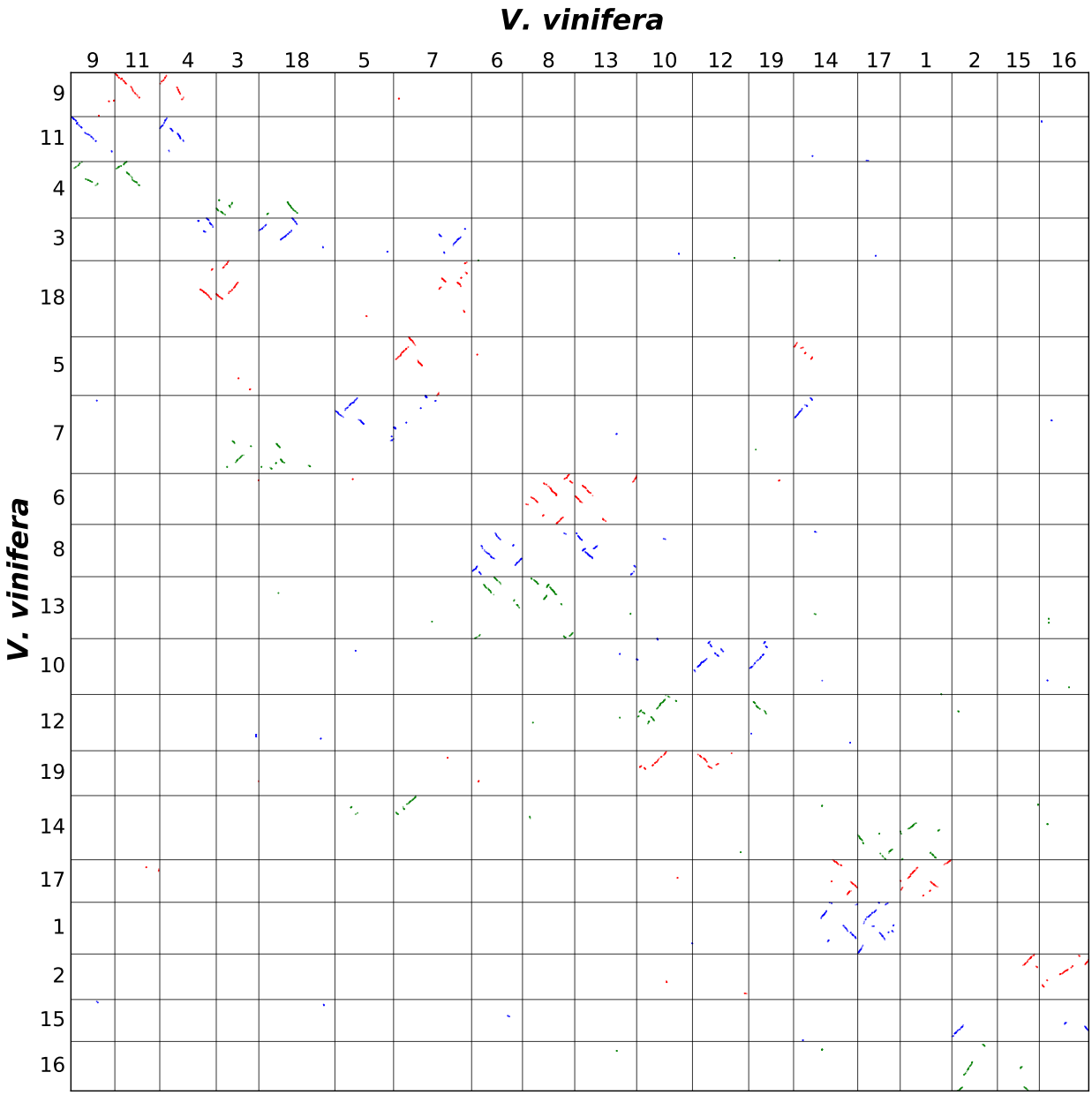

Figure S11 Collinear gene extraction of *V. vinifera*. The three candidate subsets resulted from polyploidizations were highlighted using red (S1), blue (S2), green (S3).

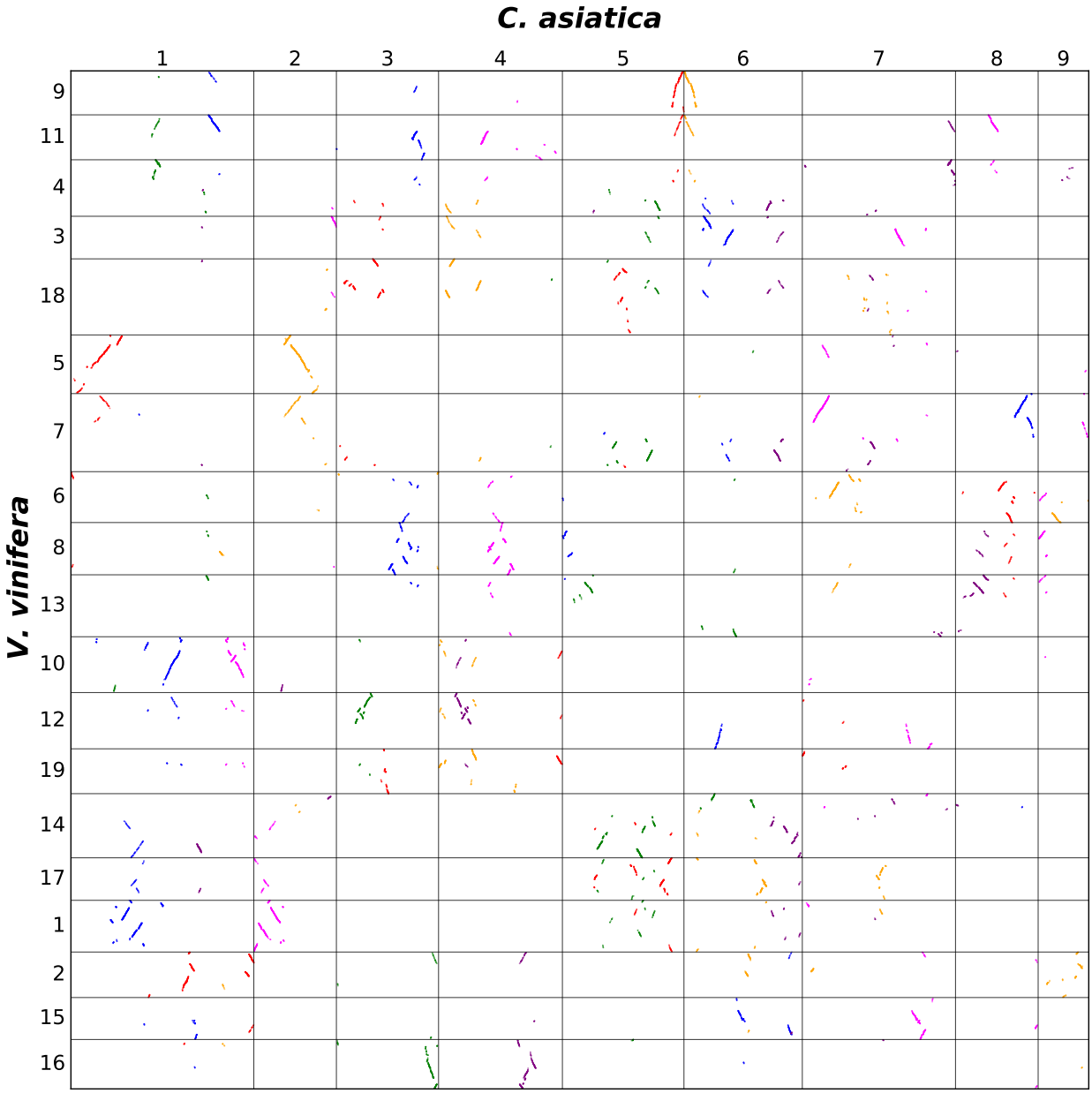

Figure S12 Collinear gene extraction between *V. vinifera* and *C. asiatica*. The six candidate subsets resulted from polyploidizations were highlighted using red (S1), blue (S2), green (S3), orange (S4), fuchsia (S5), purple (S6).

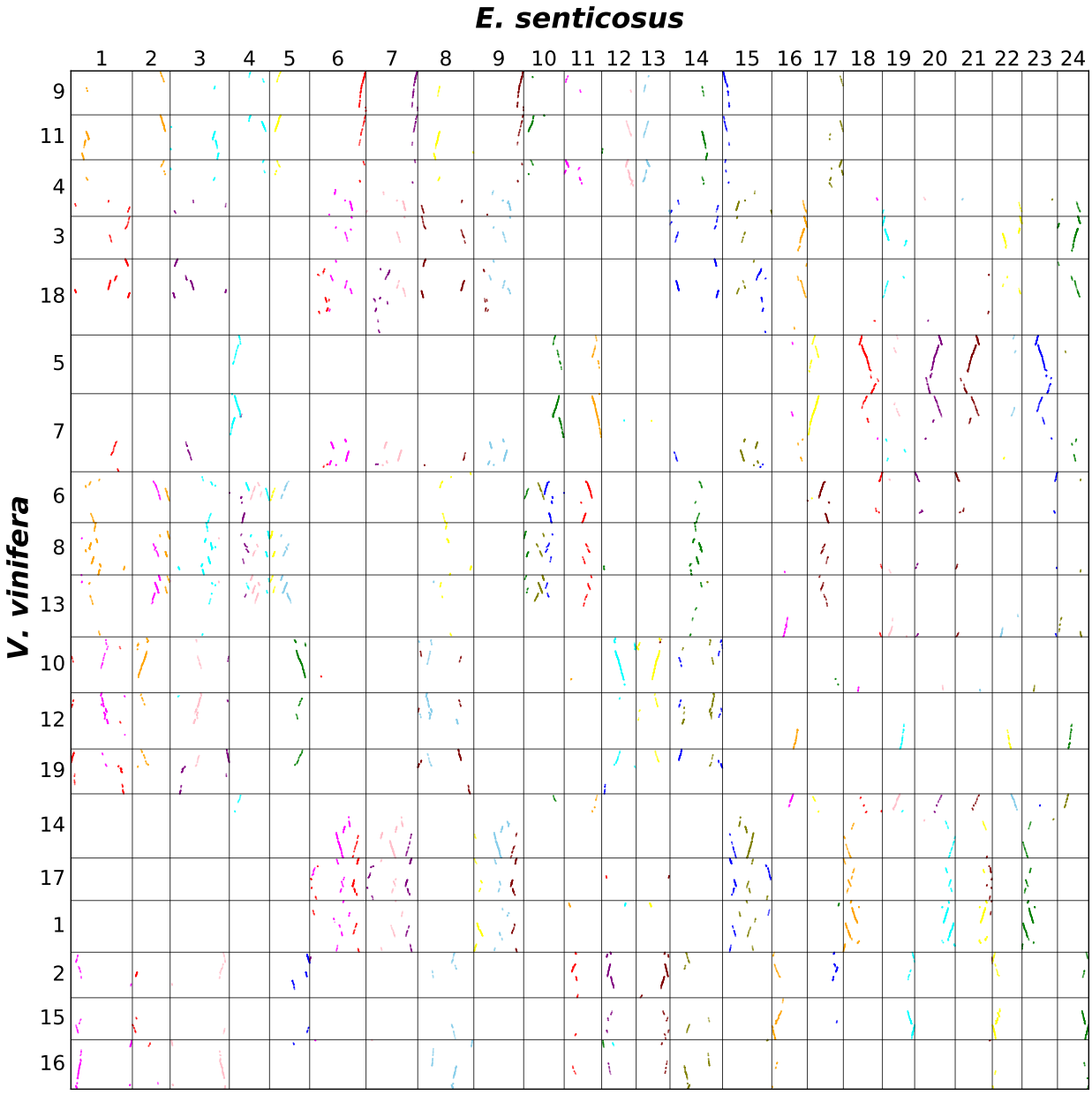

Figure S13 Collinear gene extraction between *V. vinifera* and *E. senticosus*. The 12 candidate subsets resulted from polyploidizations were highlighted using red (S1), blue (S2), green (S3), olive (S4), orange (S5), fuchsia (S6), purple (S7), cyan (S8), pink (S9), maroon (S10), yellow (S11), skyblue (S12).

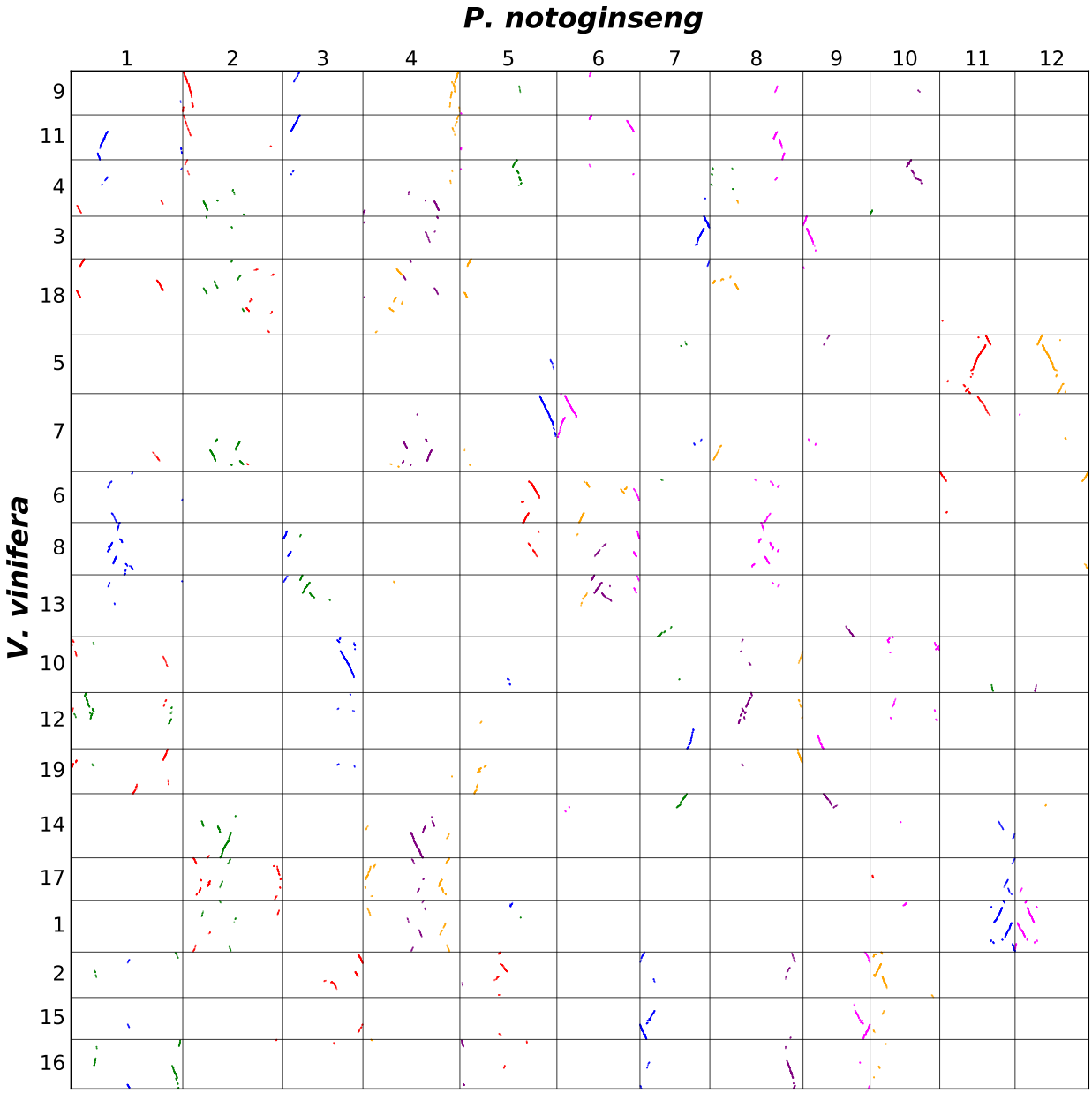

Figure S14 Collinear gene extraction between *V. vinifera* and *P. notoginseng*. The six candidate subsets resulted from polyploidizations were highlighted using red (S1), blue (S2), green (S3), orange (S4), fuchsia (S5), purple (S6).

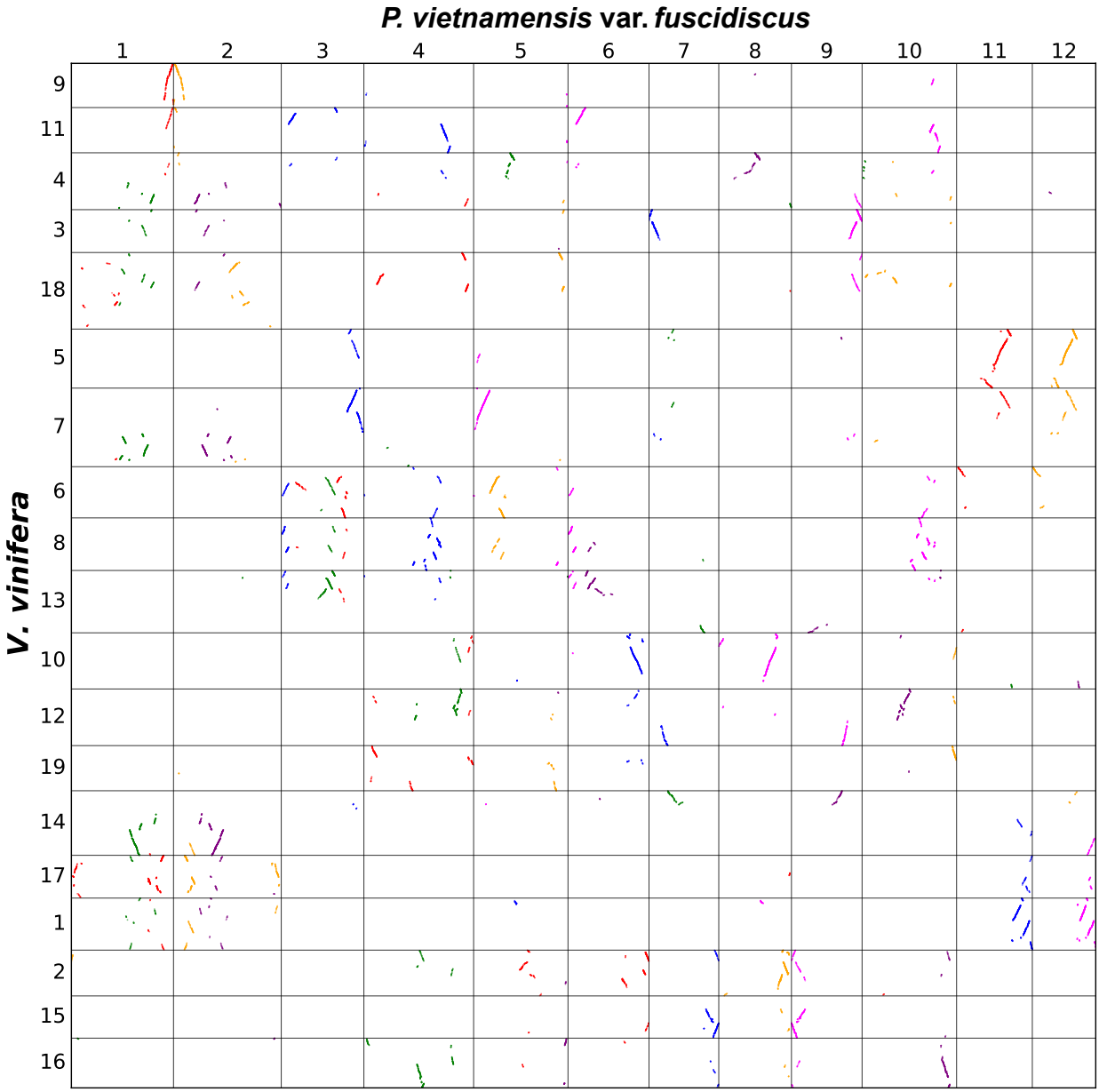

Figure S15 Collinear gene extraction between *V. vinifera* and *P. vietnamensis* var. *fuscidiscus*. The six candidate subsets resulted from polyploidizations were highlighted using red (S1), blue (S2), green (S3), orange (S4), fuchsia (S5), purple (S6).

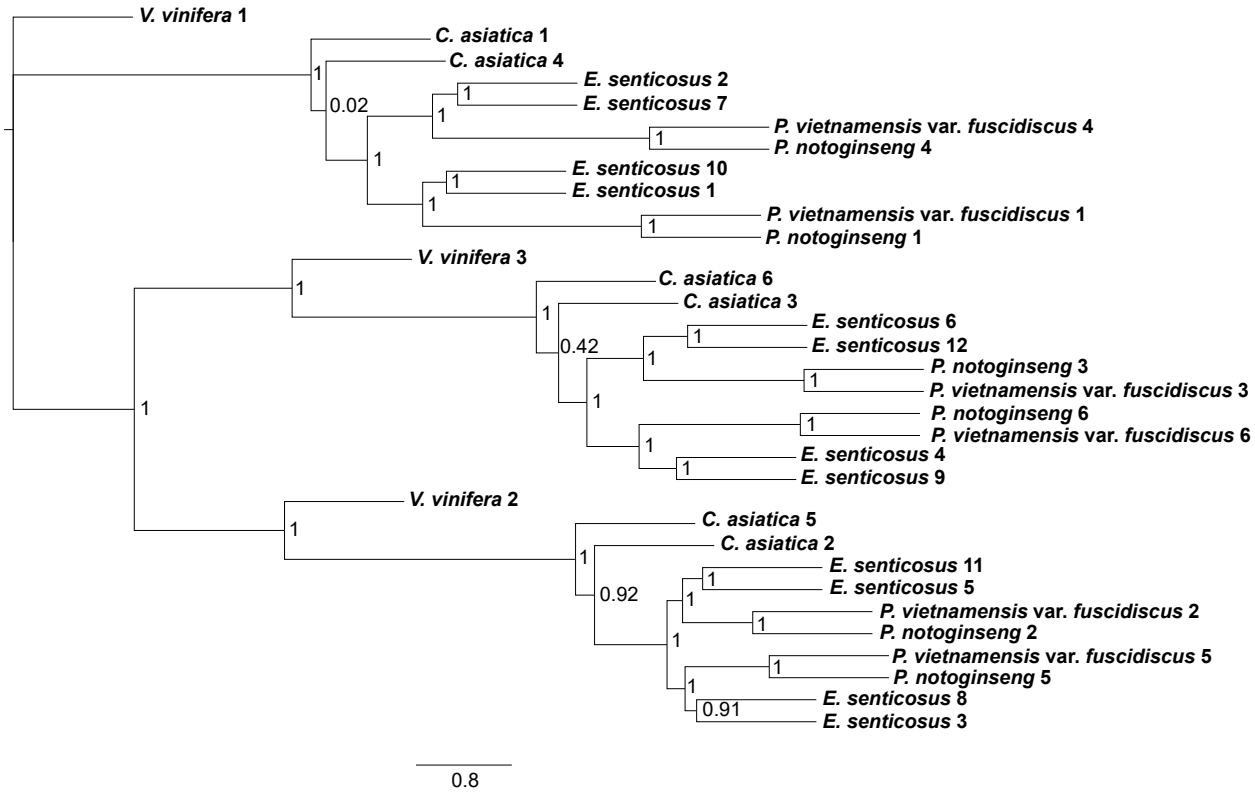

Figure S16 Synteny-based coalescent species tree including *V. vinifera* and four Apiales species. Branch lengths are shown in coalescent units. The numbers of each node represents the local posterior probabilities. Since the ASTRAL tree leaves the branch length of terminal branches empty, the length of terminal branches were all set as one.

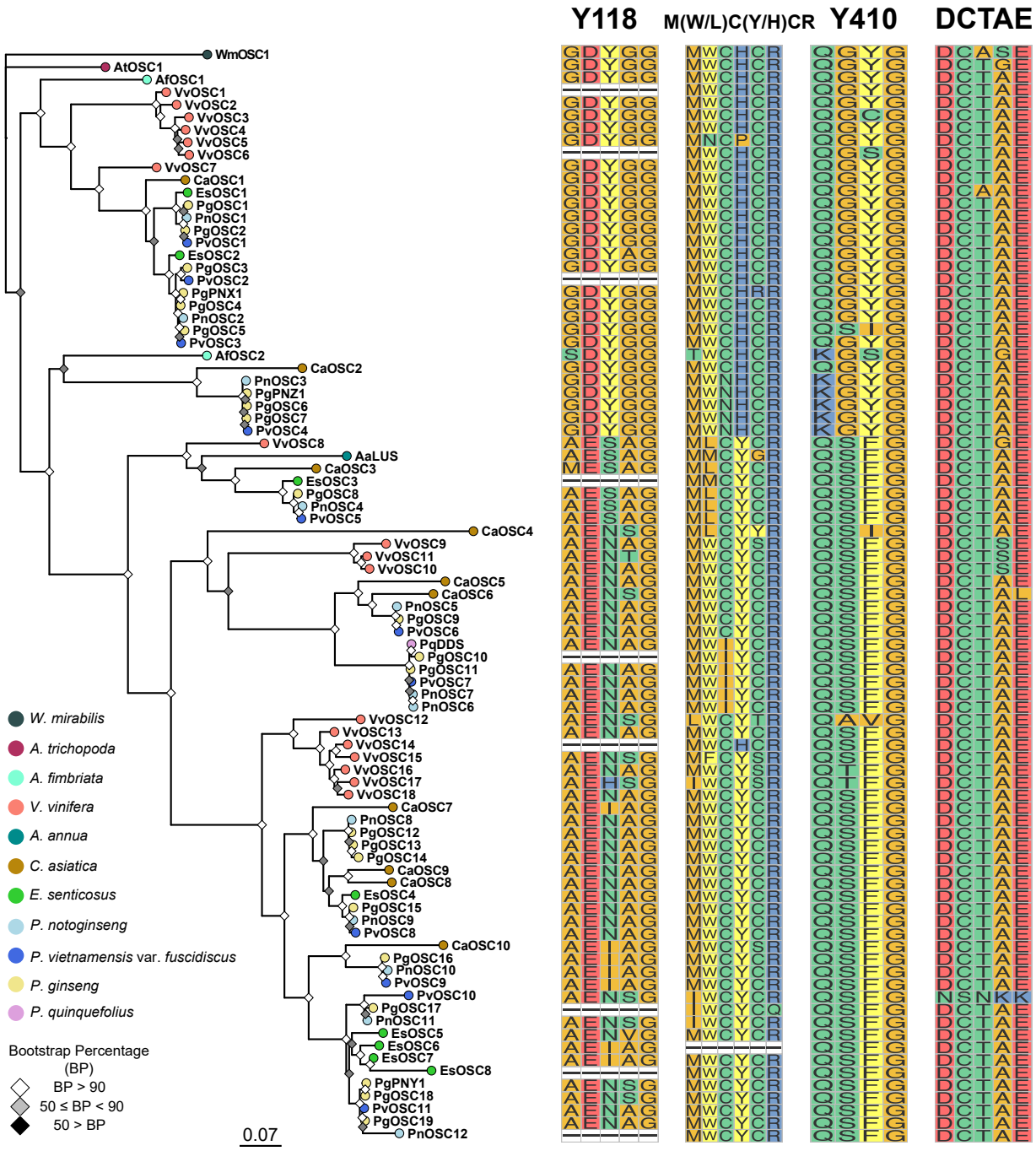

Figure S17 Maximum likelihood phylogenetic tree of OSCs with motifs aligned. Four deterministic motifs were visualized including Y118, M(W/L)C(Y/H)CR, Y410, and DCTAE.

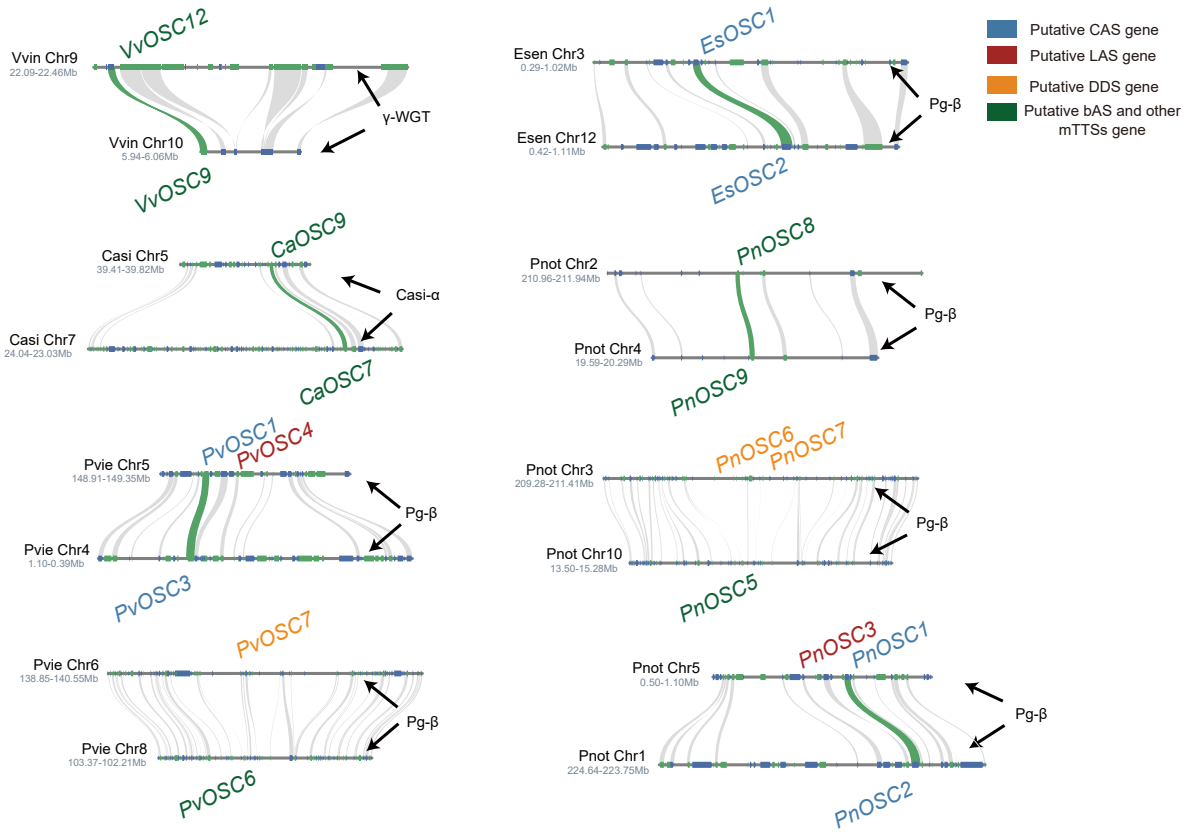

Figure S18 Intraspecific micro-synteny relations of OSC genes. Direct collinear relations for OSC genes were highlighted in green (Vvin: *V. vinifera*, Casi: *C. asiatica*, Pvie: *P. vietnamensis* var. *fuscidiscus*, Esen: *E. senticosus*, Pnot: *P. notoginseng*).

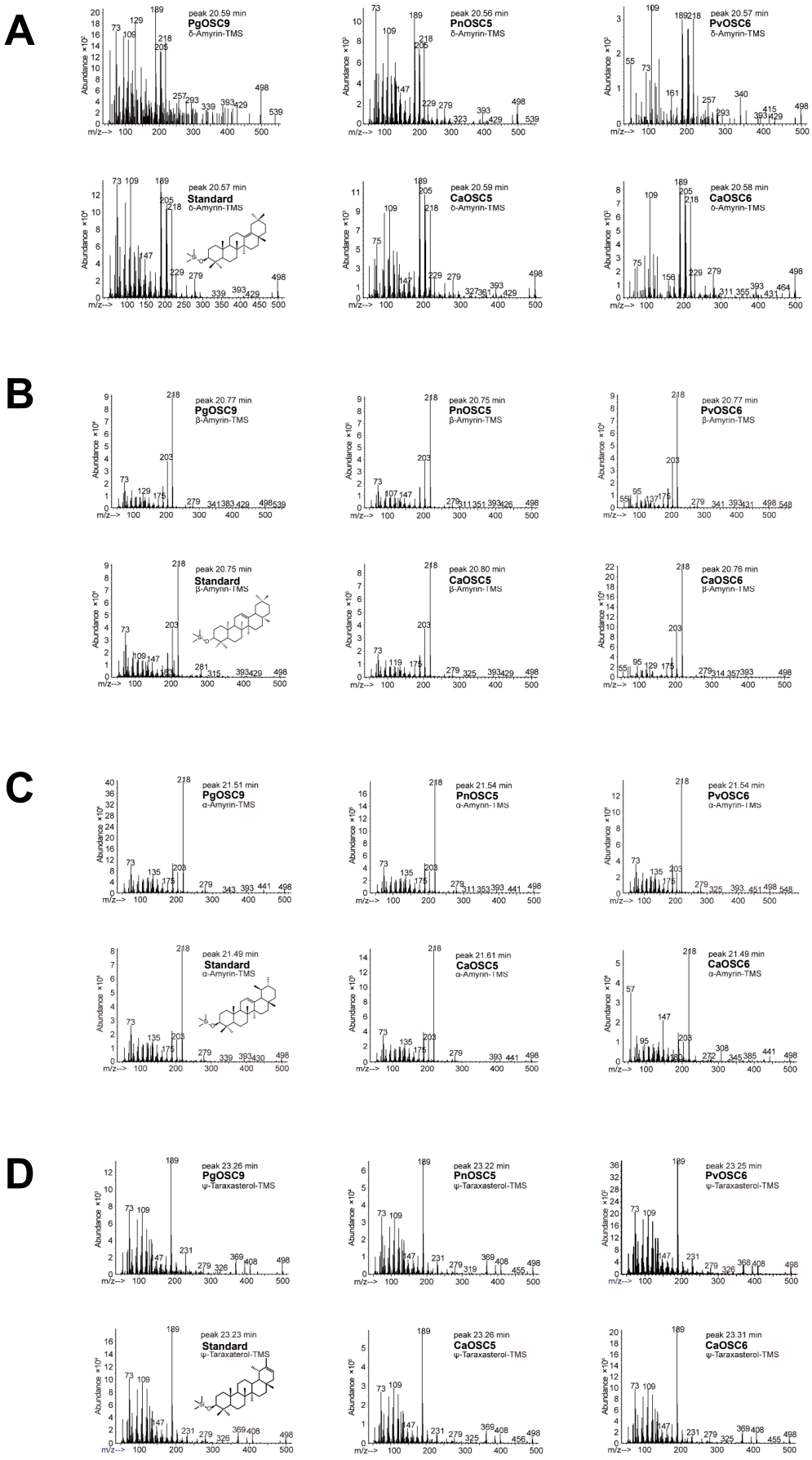

Figure S19 Mass spectra identification for compound 1 (δ-amyrin) (A), compound 2 (β-amyrin) (B), compound 3 (α-amyrin) (C), and compound 5 (ψ-taraxasterol) (D).

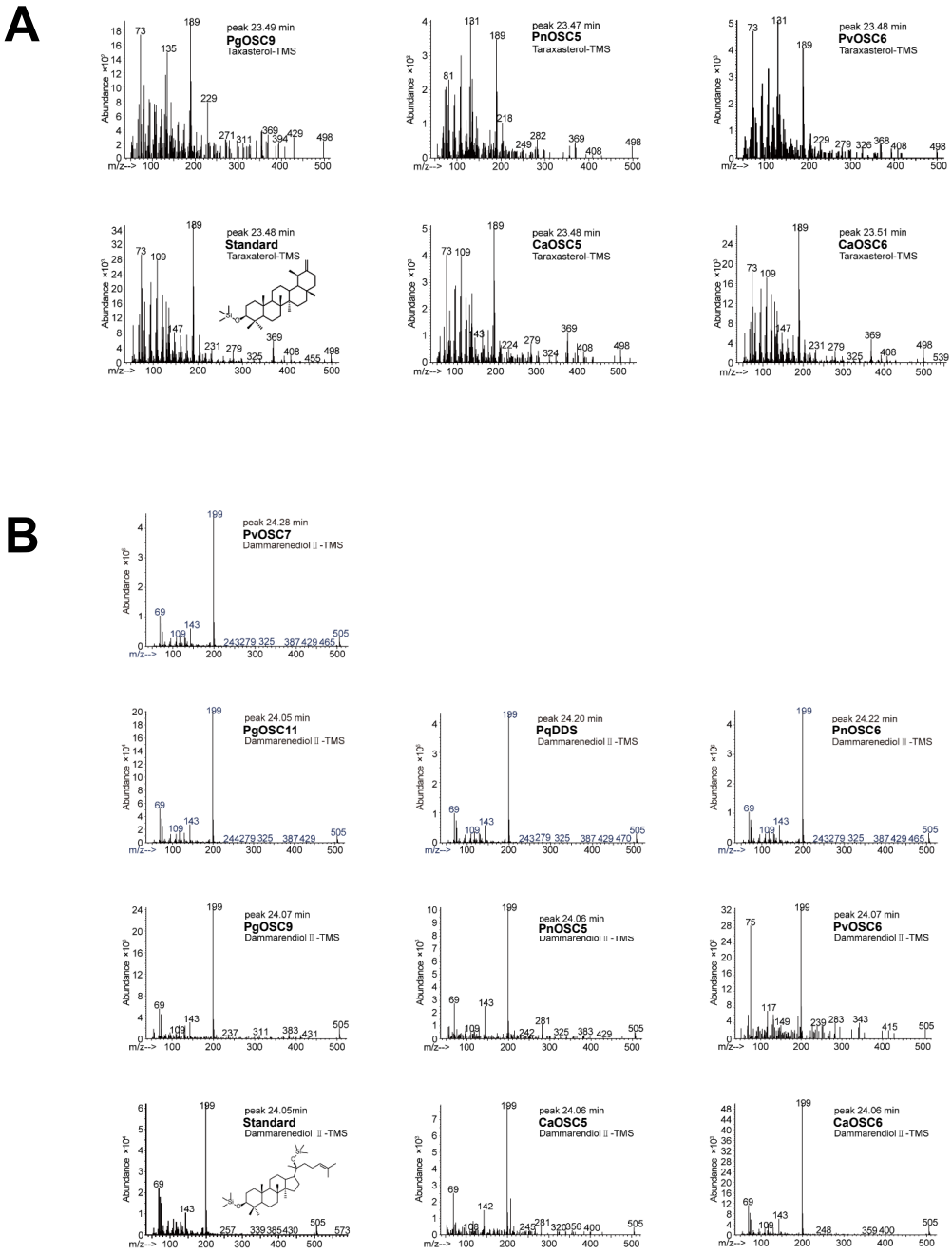

Figure S20 Mass spectra identification for compound 6 (taraxasterol) (A) and compound 7 (dammarendiol-II) (B).

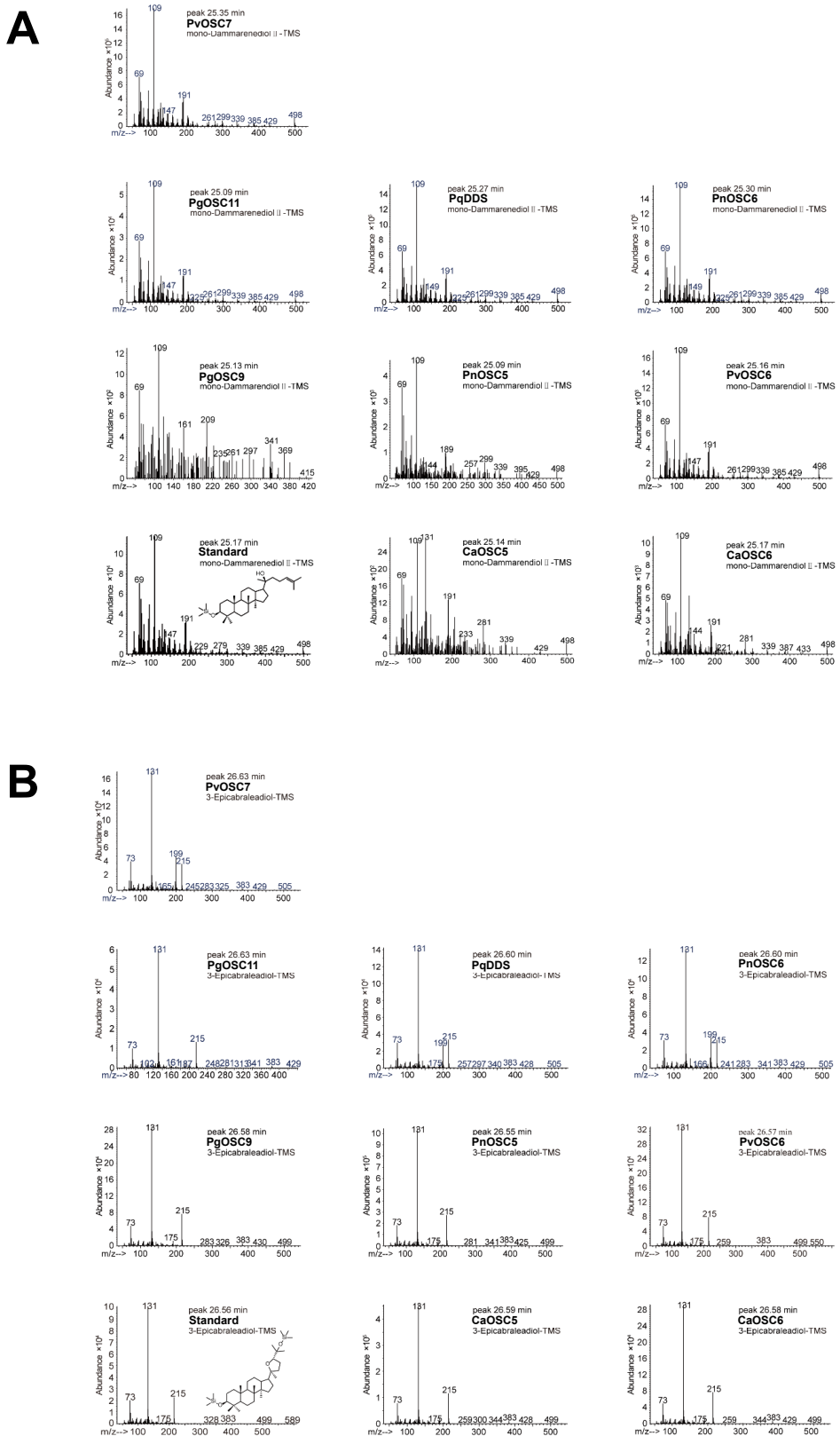

Figure S21 Mass spectra identification for compound # (A) and compound 8 (3-epicabraleadiol) (B). Compound # was identified as the mono-TMS derivative of dammarendiol-II.

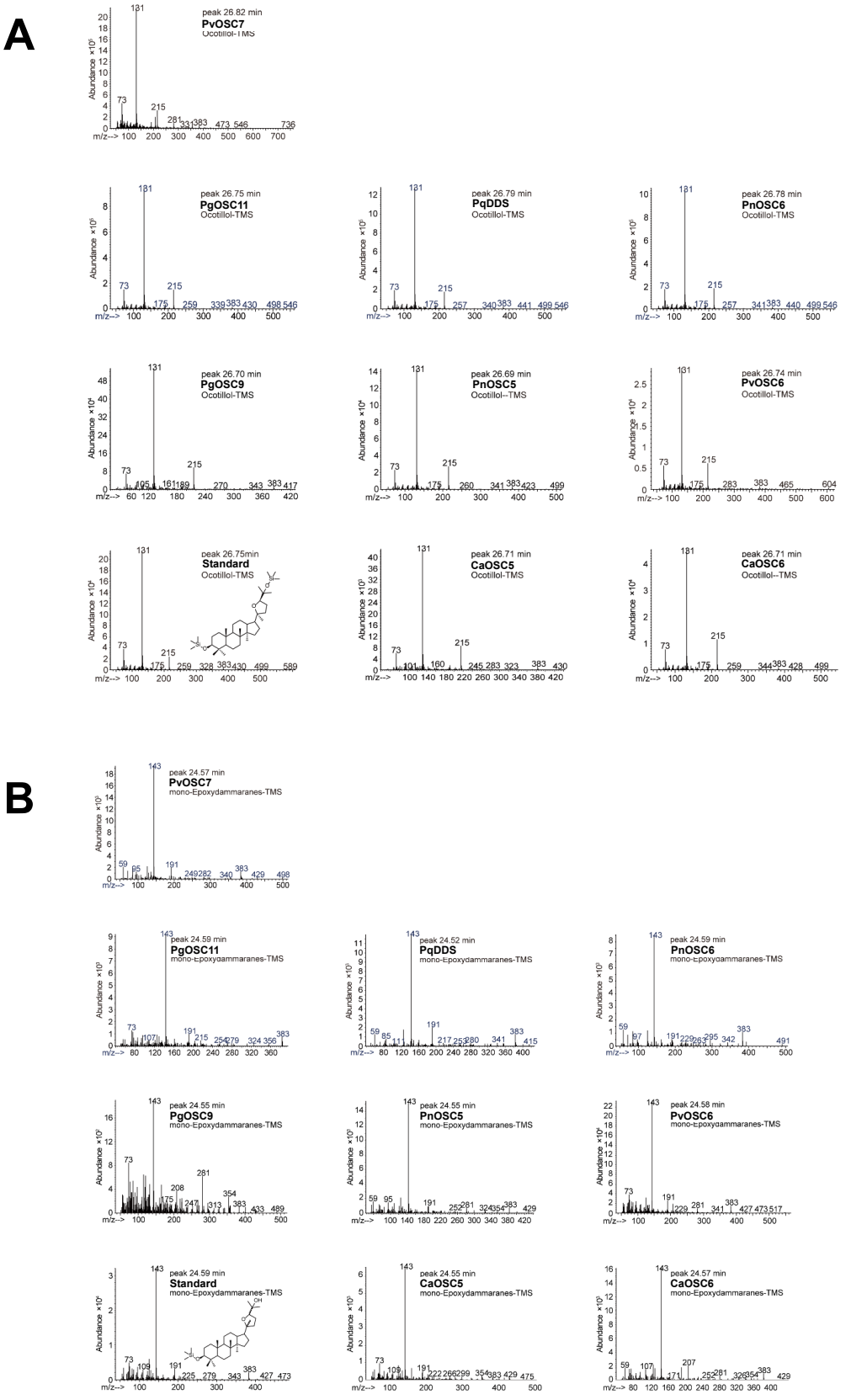

Figure S22 Mass spectra identification for compound 9 (ocotillol) (A) and compound * (B). Compound * was identified as the mono-TMS derivatives of epoxydammaranes (compound 8 and compound 9).

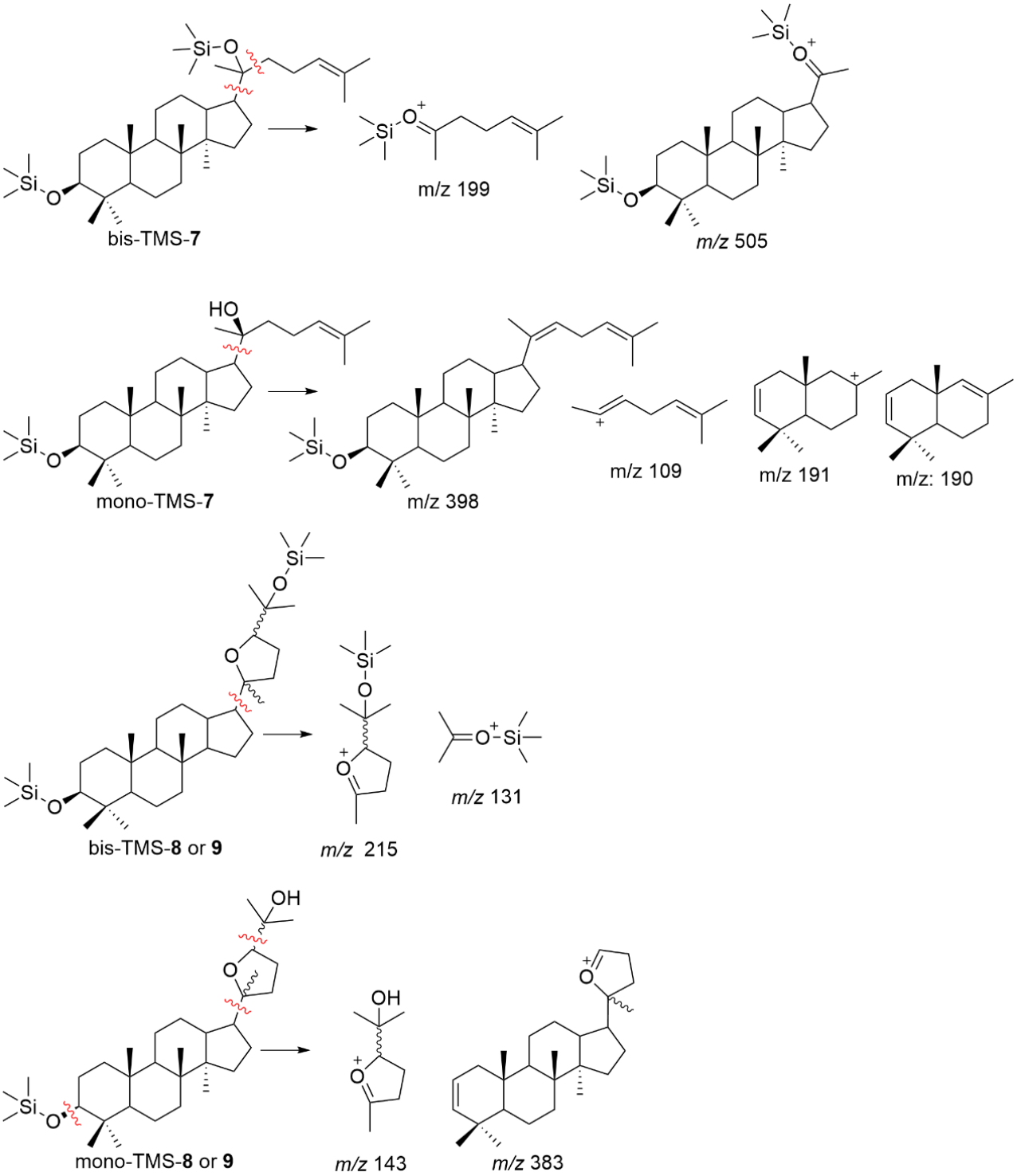

Figure S23 Proposed fragmentation patterns for compound 7, 8, and 9.

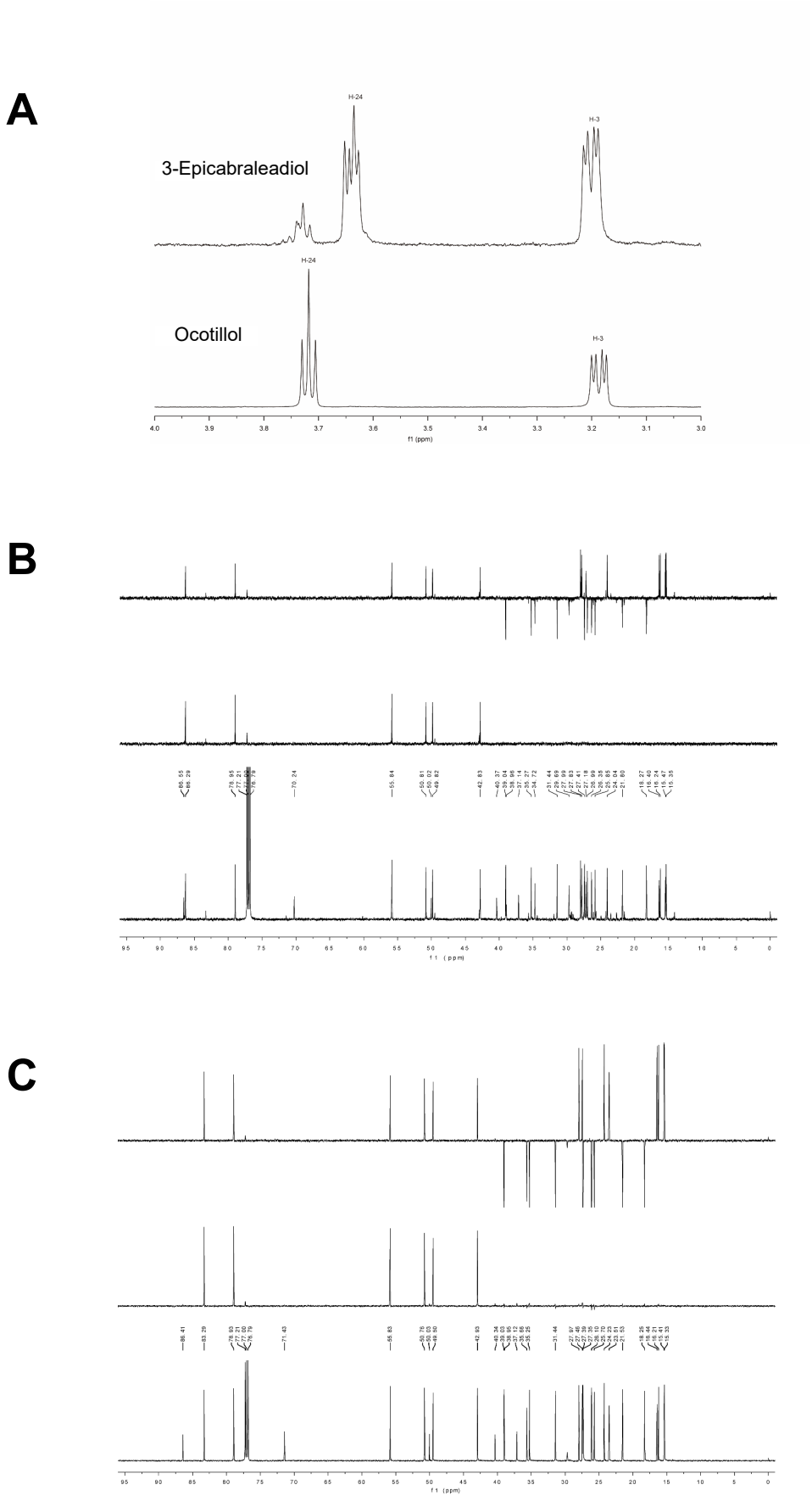

Figure S24 (A) ^{1}H NMR spectra of compound 8 (3-epicabraleadiol) and 9 (ocotillol). (B) ^{13}C NMR and DEPT spectra of compound 8 measured at 150 MHz in $CDCl_3$. (C) ^{13}C NMR and DEPT spectra of compound 9 measured at 150 MHz in $CDCl_3$.

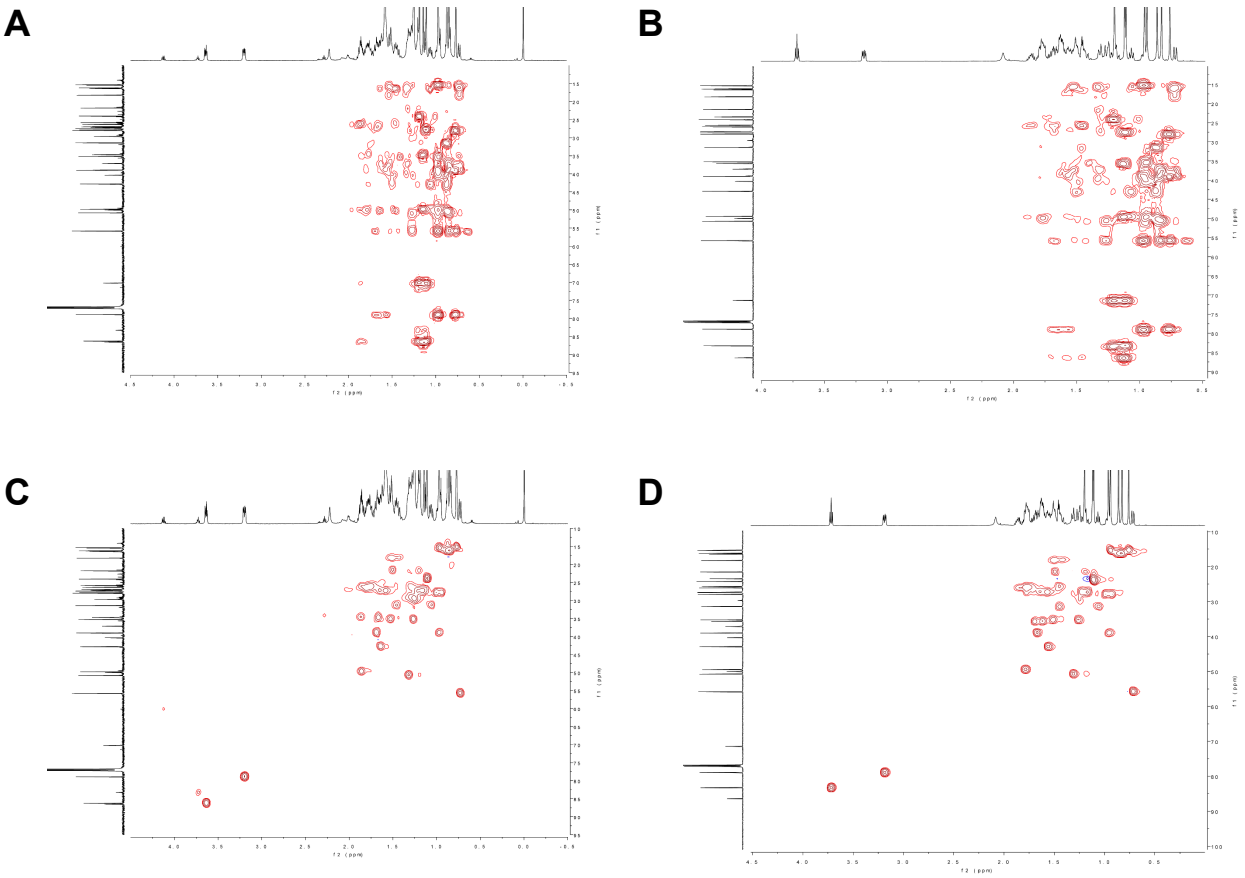

Figure S25 (A) HMBC spectrum of compound 8 measured at 600 MHz in $CDCl_3$. (B) HMBC spectrum of compound 9 measured at 600 MHz in $CDCl_3$. (C) HSQC spectrum of compound 8 measured at 600 MHz in $CDCl_3$. (D) HSQC spectrum of compound 9 measured at 600 MHz in $CDCl_3$.

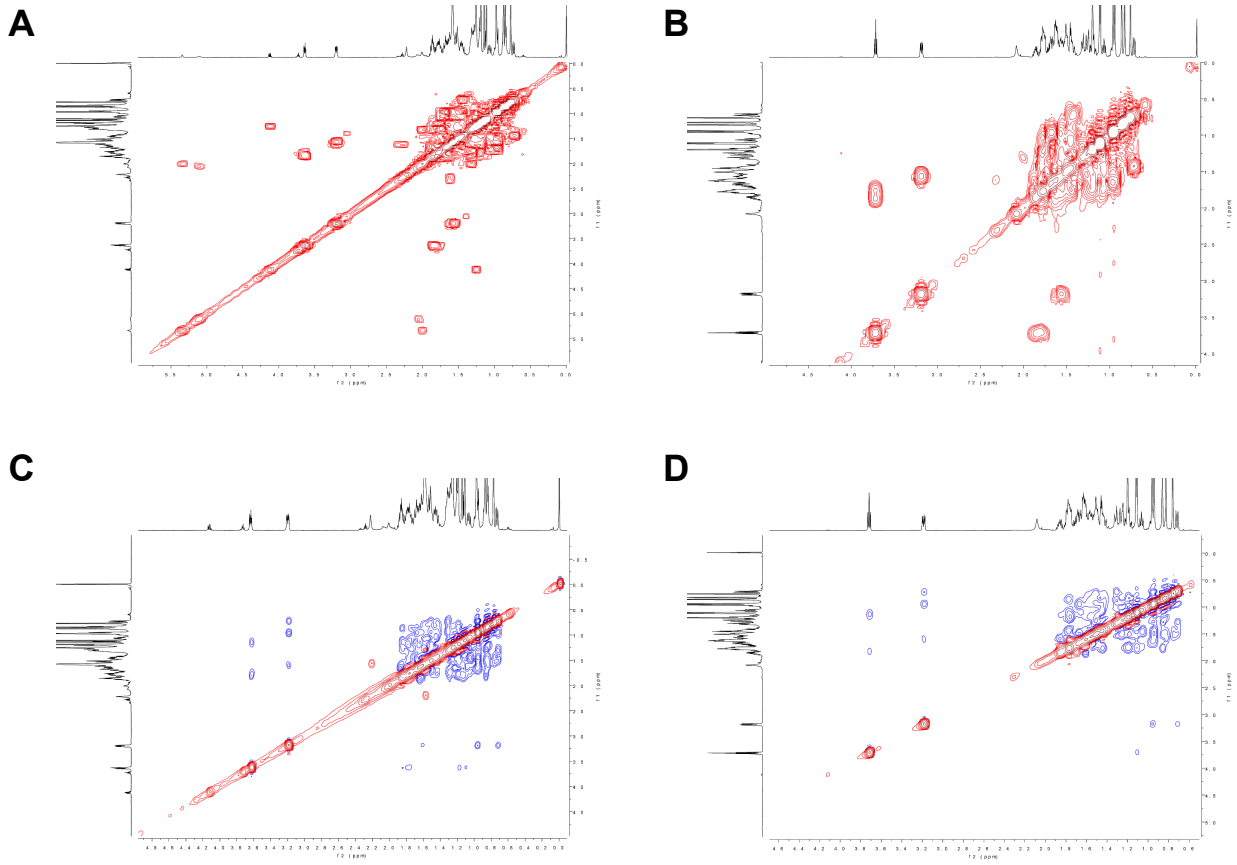

Figure S26 (A) ^{1}H-^{1}H COSY spectrum of compound 8 measured at 600 MHz in $CDCl_3$. (B) ^{1}H-^{1}H COSY spectrum of compound 9 measured at 600 MHz in $CDCl_3$. (C) ^{1}H-^{1}H ROESY spectrum of compound 8 measured at 600 MHz in $CDCl_3$. (D) ^{1}H-^{1}H ROESY spectrum of compound 9 measured at 600 MHz in $CDCl_3$.

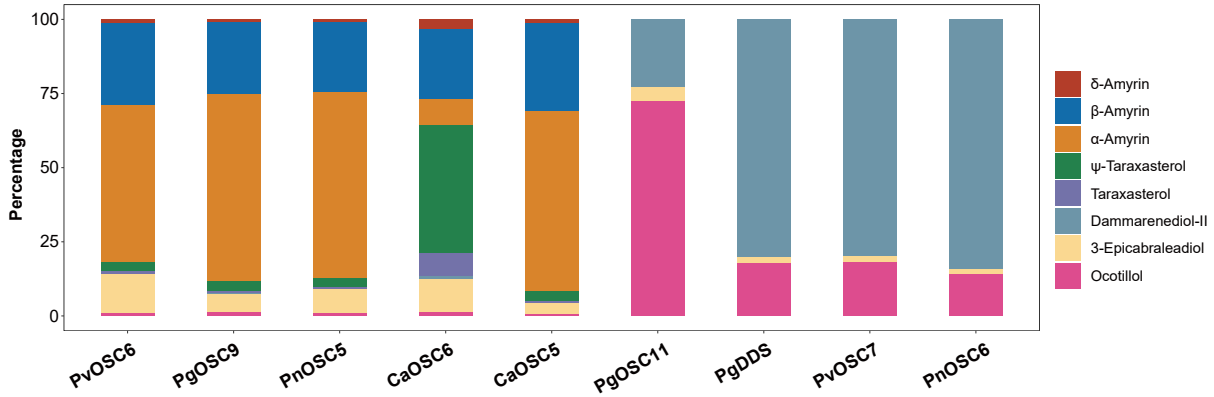

Figure S27 Relative composition of identified products for the nine OSCs. The relative abundance of each compound is calculated based on the area of the corresponding peak.

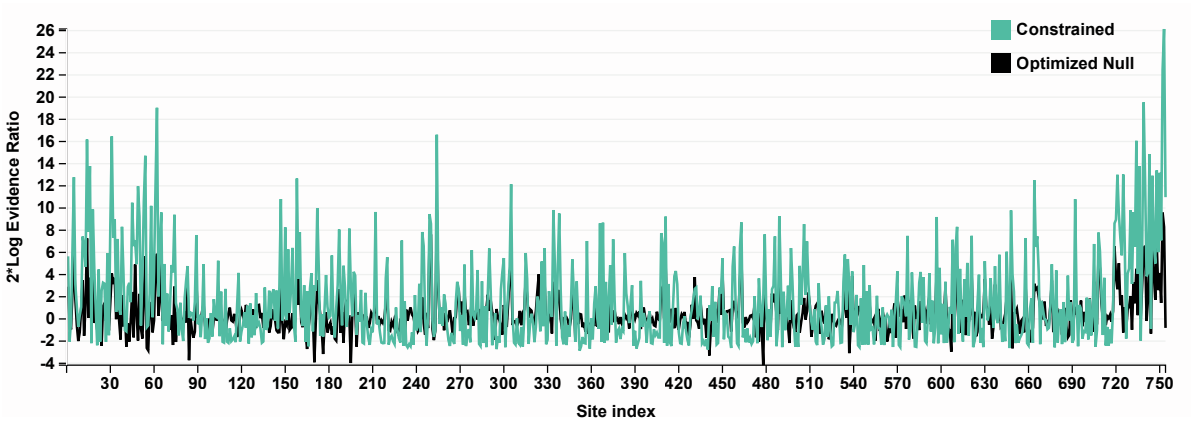

Figure S28 Evidence ratio for BUSTED model in OSCs phylogeny. BUSTED with synyonymous rate variation found evidence (LRT, P = 0.0000 $\leq$ 0.05) of gene-wide episodic diversifying selection. The Evidence ratio (y-axis) gives the likelihood ratio (on a log-scale) that the alternative model (selection along test branches) was a better fit to the data as compared to the null model.

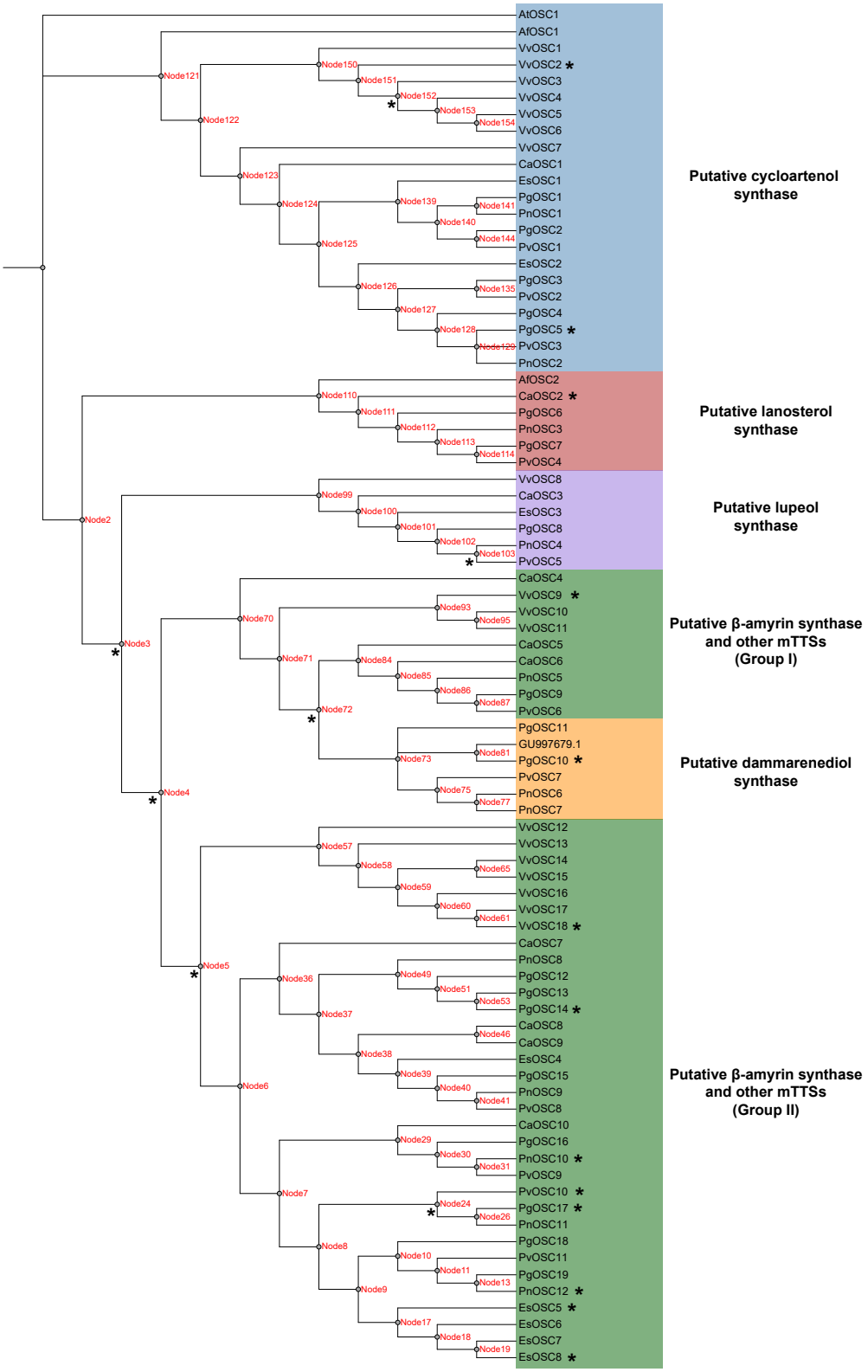

Figure S29 Maximum likelihood phylogenetic tree for OSCs showing aBSREL result for branch specific selection. The branches and internal nodes showing evidence of episodic diversifying selection were labeled with asterisks.

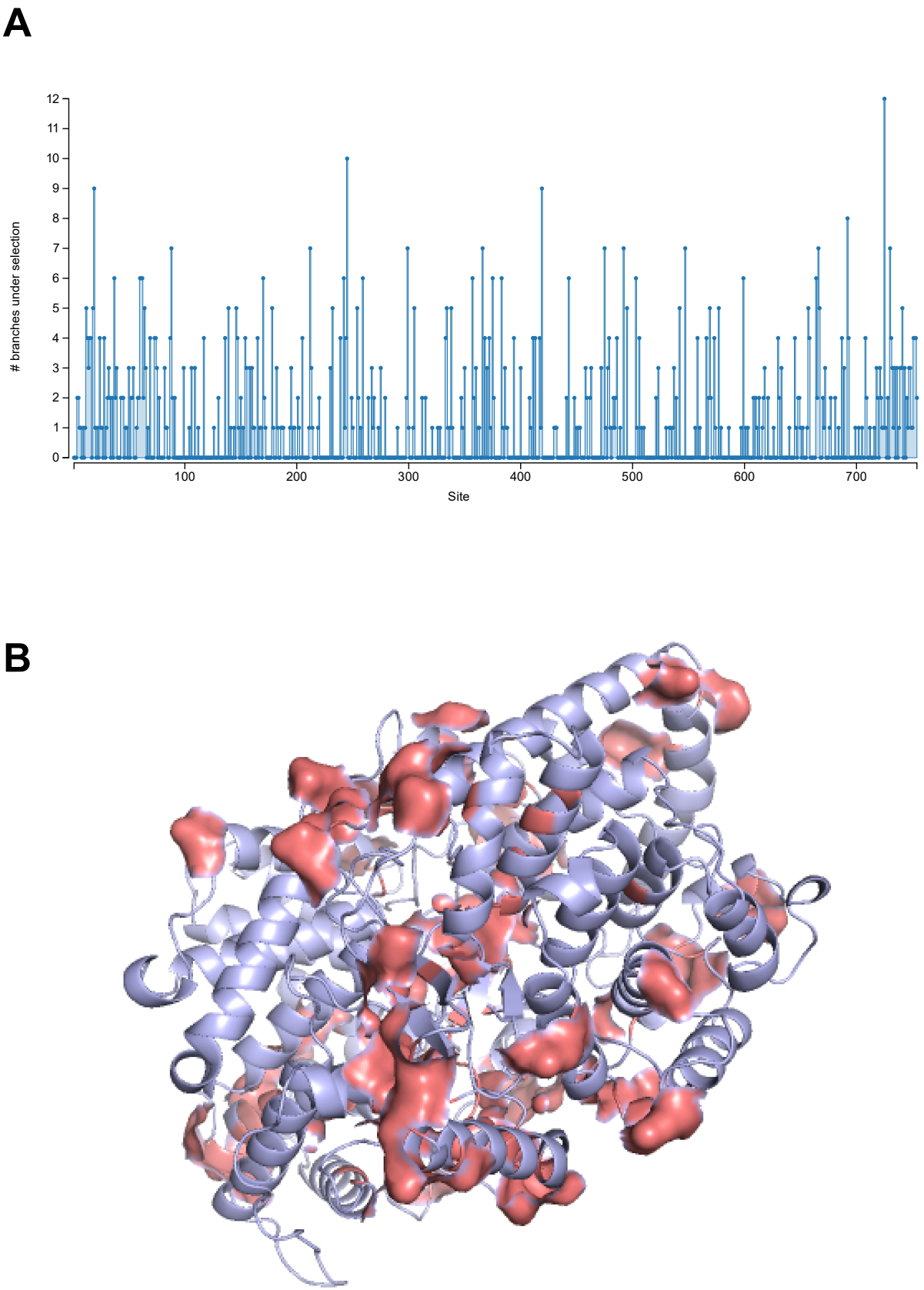

Figure S30 (A) Sites under episodic positive selection detected by MEME. Y-axis showing how many branches may have been under selection under this site (very approximate and rough). (B) 3D structure of P. ginseng CAS (AF-O82139-F1). Sites that have experienced positive selection detected by MEME were highlighted in red.

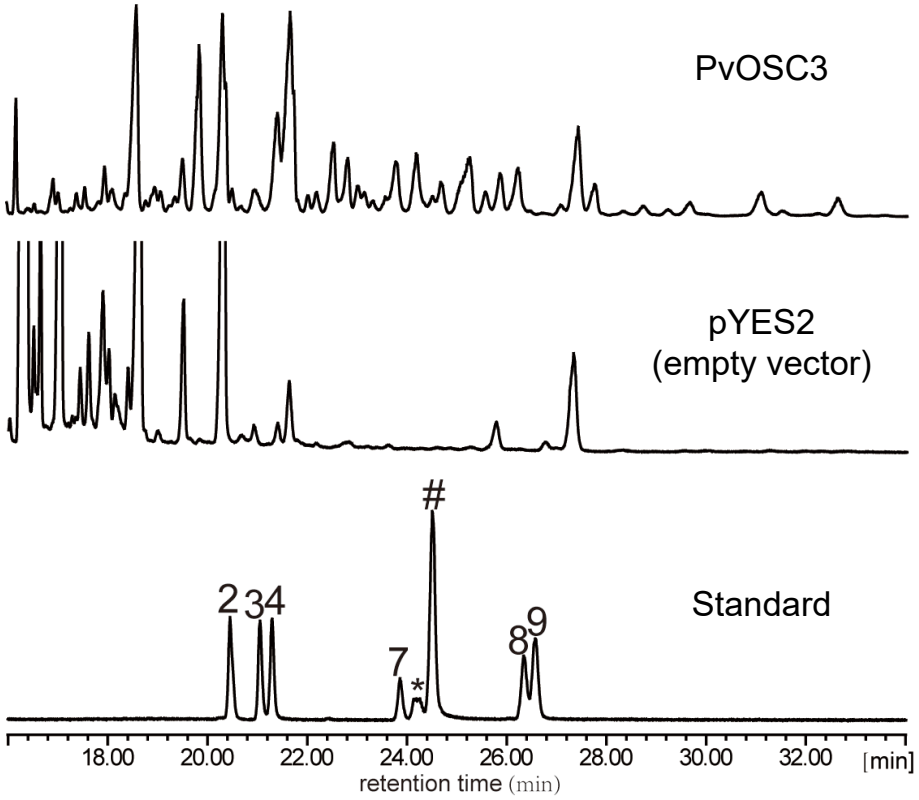

Figure S31 Functional characterization of one *P. vietnamensis* var. *fuscidiscus* CAS using heterelogous expression. * and in total ion chromatograms (TICs) represent epoxydammaranes mono-trimethylsilyl ether and dammarenediol-II mono-trimethylsilyl ether, respectively. 2: β-amyrin, 3: α-amyrin, 4: cycloartenol, 7: dammarenediol II, 8: 20S,24S-3-epicabraleadiol, 9: 20S,24R-ocotillol.

Table S1. Statistics on *P. vietnamensis* var. *fuscidiscus* genome assembly.

Item	Contig-level assembly		Hic chromosome-level assembly	
	length (bp)	Number	length (bp)	Number
N90	142,377	4,074	85,413,289	12
N80	206,622	3,082	107,381,201	10
N70	265,737	2,344	110,022,002	9
N60	334,899	1,770	120,995,942	7
N50	410,271	1,304	144,079,729	6
N40	493,050	919	149,510,626	5
N30	597,689	600	153,102,694	4
N20	730,980	339	159,980,755	3
N10	997,954	134	160,487,695	2
Max length	2,342,647	-	164,156,069	-
Total length	1,723,337,714	-	1,727,411,714	-
Total number	-	5,866	-	6,305
Average length	293,787	-	273,974	-
Number of sequences >=500bp	-	5,866	-	6,305
Number of sequences >=1000bp	-	5,866	-	6,304
Number of sequences >=2000bp	-	5,866	-	5,676
Number of sequences >=5000bp	-	5,866	-	4,664

Table S2. Genome size estimation of *P. vietnamensis* var. *fuscidiscus* using flow cytometry.

Sample ID	*Solanum lycopersicum* fluorescence intensity	*P. vietnamensis* var. *fuscidiscus* fluorescence intensity	Ratio	*S. lycopersicum* genome size (Gb)	*P. vietnamensis* var. *fuscidiscus* genome size (Gb)	Average Size (Gb)	Standard deviation
1	108.72	193.78	1.78		1.60		
2	109.86	198.5	1.81	0.90	1.63	1.61	0.01
3	110.09	196.11	1.78		1.60		

Table S3. Genome survey of *P. vietnamensis* var. *fuscidiscus* using Genomescope.

Property	min	max
Homozygous (aa)	99.14%	100%
Heterozygous (ab)	0%	0.86%
Genome Haploid Length	1,188,704,553 bp	1,430,039,526 bp
Genome Repeat Length	575,700,139 bp	692,580,802 bp
Genome Unique Length	613,004,415 bp	737,458,723 bp
Model Fit	53.47%	96.99%
Read Error Rate	0.22%	0.22%

Table S4. Mapping statistics of Illumina reads to *P. vietnamensis* var. *fuscidiscus* assembly.

	Total reads	Supplementary reads	Mapped reads	Reads paired in sequencing	Properly paired reads	Singletons (only one read mapped)	With mate mapped to a different sequence	Percentage of mapped reads	Percentage of properly paired reads	Genome coverage rate
P. vietnamensis* var. *fuscidiscus	1,546,294,209	16,501,123	1,520,692,682	1,529,793,086	1,445,135,176	3,160,273	50,057,132	98.34%	94.47%	97.40%

Table S5. Comparison of gene space of *P. vietnamensis* var. *fuscidiscus* with other species.

Species	Gene number	Average mRNA length (bp)	Total exon number	Average exon length (bp)	Average cds length per gene (bp)	Average exon number	Total intron number	Average intron length (bp)	Average intron length per gene (bp)
E. senticosus	36,372	5575.58	215,069	241.84	1429.99	5.91	180,779	834.08	4145.60
P. vietnamensis* var. *fuscidiscus	36,454	6166.47	189,971	293.88	1531.5	5.21	158,204	976.21	4236.58
P. notoginseng	36,747	6298.53	199,700	231.69	1259.06	5.43	169,123	979.01	4506.02
P. ginseng	59,352	4394.38	297,411	223.53	1120.12	5.01	241,351	760.71	3093.37
C. asiatica	27,785	3624.05	139,555	245.10	1231.04	5.02	111,770	594.88	2393.00

Table S6. Functional annotation of the predicted genes in *P. vietnamensis* var. *fuscidiscus*.

Type		Number	Percentage (%)
Total		36,454	100
Annotated	**eggNOG**	33,570	92.09
	GO	17,491	47.98
	KEGG	16,023	43.95
	PFAM domains	31,444	86.26
Unannotated		2,884	7.91

Table S7. BUSCO analysis results of genome assemblies and annotated gene sets for *P. vietnamensis* var. *fuscidiscus* and *P. notoginseng*.

Species	Items	Complete BUSCOs	Complete and single-copy BUSCOs	Complete and duplicated BUSCOs	Fragmented BUSCOs	Missing BUSCOs	Total BUSCO groups searched
P. vietnamensis* var. *fuscidiscus	Genome (full assembly)	95.3%	84.6%	10.7%	1.0%	3.7%	2,326
	Genome (pseudo chromosomes)	95.0%	84.9%	10.1%	1.0%	4.0%	2,326
	Gene set	92.6%	83.4%	9.2%	2.2%	5.2%	2,326
P. notoginseng	Genome (full assembly)	97.5%	84.2%	13.3%	1.0%	1.5%	2,326
	Genome (pseudo chromosomes)	96.8%	85.5%	11.3%	1.2%	2.0%	2,327
	Gene set	93.3%	81.7%	11.6%	1.9%	4.8%	2,326

Table S8. Statistics on *P. notoginseng* genome assembly (updated).

Item	length (bp)	Number
N90	152,644,078	12
N80	161,973,221	10
N70	169,268,257	9
N60	176,594,900	7
N50	196,656,904	6
N40	200,662,011	5
N30	213,721,665	4
N20	216,551,805	3
N10	221,362,758	2
Max length	225,175,661	-
Total length	2,402,896,139	-
Total number	-	5,223
Average length	460,060	-
Number of sequences >=500bp	-	5,223
Number of sequences >=1000bp	-	5,223
Number of sequences >=2000bp	-	4,755
Number of sequences >=5000bp	-	4,097

Table S9. Functional annotation of the predicted genes in *P. notoginseng*.

Type		Number	Percentage (%)
Total		36,747	100
Annotated	**eggNOG**	34,098	92.79
	GO	18,151	49.39
	KEGG	17,214	46.84
	PFAM domains	30,993	84.34
Unannotated		2,649	7.21

Table S10. Repeat annotation of *P. vietnamensis* var. *fuscidiscus*.

Type		Length (bp)	Percentage of genome (%)
Tandem repeats	**Simple repeats**	101,728,874	5.90
	Satellite repeats	64,151	0.00
Interspersed repeats	**LTR**	1,360,330,196	78.94
	LTR (Copia)	97,780,472	5.67
	LTR (Gypsy)	939,613,786	54.52
	SINE	39,882	0.00
	LINE	9,521,199	0.55
	DNA transposons	50,052,051	2.90
	Unclassified	100,930,458	5.86
Total (non-redundant)		1,495,684,042	86.79
LTR identity		95.26	
raw LTR assembly index (LAI)		15.18	
LAI		11.63	

Table S11. Repeat annotation of *P. notoginseng*.

Type		Length (bp)	Percentage of genome (%)
Tandem repeats	**Simple repeats**	135,431,446	5.64
	Satellite repeats	1,939	0.00
Interspersed repeats	**LTR**	1,938,176,020	80.66
	LTR (Copia)	109,347,216	4.55
	LTR (Gypsy)	1,333,339,204	55.49
	SINE	8,342	0.00
	LINE	9,181,808	0.38
	DNA transposons	75,167,439	3.13
	Unclassified	134,389,290	5.59
Total (non-redundant)		2,118,941,998	88.18
LTR identity			94.33
raw LTR assembly index (LAI)			11.87
LAI			10.95

Table S12. Classification of transposable elements in *P. vietnamensis* var. *fuscidiscus*.

Class	Order	Superfamily	Clade	Number
Class I retrotransposons	**LTR**	**Copia**	**Ale**	3,358
			Alesia	111
			Angela	7,560
			Bianca	2,232
			Bryco	17
			Lyco	17
			Gymco-III	5
			Gymco-I	4
			Gymco-II	22
			Ikeros	1,537
			Ivana	950
			Gymco-IV	24
			Osser	16
			SIRE	12,000
			TAR	2,070
			Tork	1,487
			mixture/unknown	98,537
		Gypsy	**non-chromo-outgroup**	16
			Phygy	3
			Selgy	3
			Athila	12,422
			TatI	8
			TatII	126
			TatIII	59
			Ogre	19,942
			Retand	7,292
			Chlamyvir	82

				Tcn1	18
				chromo-outgroup	64
				CRM	4,253
				Galadriel	167
				Tekay	233,307
				Reina	3,261
				chromo-unclass	8
				mixture/unknown	639,713
		Retrovirus	**unknown**	**unknown**	2,145
		pararetrovirus	**unknown**	**unknown**	3,263
		DIRS	**unknown**	**unknown**	316
		LINE	**unknown**	**unknown**	6,028
Class II DNA	**Subclass**	**TIR**	**EnSpm_CACTA**	**unknown**	3,752
transposons	**1**		**hAT**	**unknown**	3,322
			Merlin	**unknown**	1,762
			MuDR_Mutator	**unknown**	2,708
			PIF_Harbinger	**unknown**	482
			Sola1	**unknown**	1
			Tc1_Mariner	**unknown**	280
	Subclass	**Helitron**	**unknown**	**unknown**	667
	2	**Maverick**	**unknown**	**unknown**	4,085
mixture		**mixture**	**mixture**	**unknown**	377

Table S13. Classification of transposable elements in *P. notoginseng*.

Class	Order	Superfamily	Clade	Number
Class I retrotransposons	**LTR**	**Copia**	**Ale**	4,060
			Alesia	155
			Angela	15,715
			Bianca	2,933
			Bryco	13
			Lyco	13
			Gymco-III	6
			Gymco-I	15
			Gymco-II	22
			Ikeros	1,820
			Ivana	1,389
			Gymco-IV	16
			Osser	14
			SIRE	10,494
			TAR	2,481
			Tork	1,708
			mixture/unknown	114,324
		Gypsy	**non-chromo-outgroup**	29
			Phygy	4
			Selgy	3
			Athila	18,508
			TatI	7
			TatII	236
			TatIII	75
			Ogre	34,678
			Retand	15,239
			Chlamyvir	99

				Tcn1	16
				chromo-outgroup	57
				CRM	4,791
				Galadriel	232
				Tekay	332,594
				Reina	4,502
				chromo-unclass	22
				mixture/unknown	999,717
		Retrovirus	**unknown**	**unknown**	584
		pararetrovirus	**unknown**	**unknown**	4,309
		DIRS	**unknown**	**unknown**	1,226
		LINE	**unknown**	**unknown**	3,373
Class II DNA	**Subclass**	**TIR**	**EnSpm_CACTA**	**unknown**	5,060
transposons	**1**		**hAT**	**unknown**	3,539
			Merlin	**unknown**	1,693
			MuDR_Mutator	**unknown**	5,574
			PIF_Harbinger	**unknown**	1,059
			Sola1	**unknown**	18
			Tc1_Mariner	**unknown**	1,090
	Subclass	**Helitron**	**unknown**	**unknown**	1,769
	2	**Maverick**	**unknown**	**unknown**	3,616
mixture		**mixture**	**mixture**	**unknown**	560

Table S14. Summary of gene family analysis of *P. vietnamensis* var. *fuscidiscus* with other species.

Type	Number
Number of species	12
Number of genes	416,694
Number of genes in orthogroups	387,841
Number of unassigned genes	28,853
Percentage of genes in orthogroups	93.1
Percentage of unassigned genes	6.9
Number of orthogroups	30,074
Number of species-specific orthogroups	7,487
Number of genes in species-specific orthogroups	36,109
Percentage of genes in species-specific orthogroups	8.7
Mean orthogroup size	12.9
Median orthogroup size	10
G50 (assigned genes)	20
G50 (all genes)	19
O50 (assigned genes)	5,736
O50 (all genes)	6,486
Number of orthogroups with all species present	8,542
Number of single-copy orthogroups	168

Table S15. Results of GO enrichment analysis of expanded gene families in *P. vietnamensis* var. *fuscidiscus* ($P < 0.05$).

ID	Description	GeneRatio	BgRatio	p.adjust	qvalue	Count
GO:0030246	carbohydrate binding	40/253	406/17485	3.37325E-20	2.24067E-20	40
GO:0004553	hydrolase activity, hydrolyzing O-glycosyl compounds	38/253	410/17485	3.15485E-18	2.0956E-18	38
GO:0016798	hydrolase activity, acting on glycosyl bonds	38/253	446/17485	4.66915E-17	3.10146E-17	38
GO:0004565	beta-galactosidase activity	32/253	106/17485	4.15053E-31	2.75698E-31	32
GO:0015925	galactosidase activity	32/253	114/17485	3.86914E-30	2.57007E-30	32
GO:0120251	hydrocarbon biosynthetic process	27/253	68/17485	4.60183E-30	3.05675E-30	27
GO:0120252	hydrocarbon metabolic process	27/253	77/17485	1.17591E-28	7.81096E-29	27
GO:0000287	magnesium ion binding	24/253	224/17485	9.14711E-13	6.07594E-13	24
GO:0010333	terpene synthase activity	22/253	32/17485	4.15053E-31	2.75698E-31	22
GO:0016838	carbon-oxygen lyase activity, acting on phosphates	22/253	37/17485	1.80523E-29	1.19912E-29	22
GO:0016835	carbon-oxygen lyase activity	22/253	167/17485	1.69374E-13	1.12506E-13	22
GO:0010334	sesquiterpene synthase activity	19/253	24/17485	2.4489E-29	1.62668E-29	19
GO:0051761	sesquiterpene metabolic process	19/253	24/17485	2.4489E-29	1.62668E-29	19
GO:0051762	sesquiterpene biosynthetic process	19/253	24/17485	2.4489E-29	1.62668E-29	19
GO:0034247	snoRNA splicing	10/253	10/17485	1.95372E-17	1.29776E-17	10

GO:0080013	(E,E)-geranyllinalool synthase activity	10/253	10/17485	1.95372E-17	1.29776E-17	10
GO:0010623	programmed cell death involved in cell development	10/253	22/17485	7.74677E-12	5.14577E-12	10
GO:0045292	mRNA cis splicing, via spliceosome	10/253	23/17485	1.23559E-11	8.20739E-12	10
GO:0080027	response to herbivore	9/253	16/17485	9.99025E-12	6.63599E-12	9
GO:0016635	oxidoreductase activity, acting on the CH-CH group of donors, quinone or related compound as acceptor	8/253	13/17485	7.23975E-11	4.80898E-11	8

Table S16. Results of KEGG enrichment analysis of expanded gene families in *P. vietnamensis* var. *fuscidiscus* ($P < 0.05$).

Term Name	MainClass	GeneRatio	BgRatio	enrichFactor	corrected p-value(BH method)
B 09131 Membrane transport	A09130 Environmental Information Processing	17/331	90/15326	8.745955	1.993E-10
02010 ABC transporters	A09130 Environmental Information Processing	17/331	90/15326	8.745955	1.993E-10
04090 CD molecules	A09180 Brite Hierarchies	10/331	65/15326	7.1234023	1.427E-05
00909 Sesquiterpenoid and triterpenoid biosynthesis	A09100 Metabolism	25/331	201/15326	5.7589695	6.421E-11
00904 Diterpenoid biosynthesis	A09100 Metabolism	5/331	42/15326	5.5121565	0.015191
03032 DNA replication proteins	A09180 Brite Hierarchies	20/331	240/15326	3.8585096	3.761E-06
00240 Pyrimidine metabolism	A09100 Metabolism	8/331	101/15326	3.6674942	0.0127851
04120 Ubiquitin mediated proteolysis	A09120 Genetic Information Processing	15/331	213/15326	3.2607123	0.0006457
B 09109 Metabolism of terpenoids and polyketides	A09100 Metabolism	31/331	498/15326	2.8822602	2.124E-06
03036 Chromosome and associated proteins	A09180 Brite Hierarchies	64/331	1131/15326	2.620102	5.7E-11
A09130 Environmental Information Processing	A09130 Environmental Information Processing	30/331	652/15326	2.1304654	0.0007545

Table S17. Summary of Gaussian kernel analysis on Ks distribution of intraspecific and interspecific colinear gene blocks.

Intraspecific/interspecific colinear gene blocks	Peak of Ks distribution	Deviation	R square of linear regression
A. graveolens (gamma)	1.793	0.1824	0.8052
A. graveolens (Apiaceae-β)	1.0479	0.1343	0.9315
A. graveolens (Apiaceae-α)	0.5749	0.086	0.9434
C. asiatica (gamma)	1.7442	0.2825	0.9207
C. asiatica （Casi-α)	0.7481	0.1003	0.9513
E. senticosus (gamma)	1.4778	0.1737	0.9262
E. senticosus (Pg-β)	0.3818	0.0338	0.9718
E. senticosus (Esen-α)	0.137	0.021	0.9968
P. notoginseng (gamma)	1.5018	0.2433	0.9272
P. notoginseng (Pg-β)	0.3837	0.0362	0.9312
P. vietnamensis var. *fuscidiscus* (gamma)	1.5116	0.2141	0.9418
P. vietnamensis var. *fuscidiscus* (Pg-β)	0.3773	0.0502	0.9604
V. vinifera (gamma)	1.2852	0.155	0.9624
A. graveolens-*V. vinifera* (gamma)	1.6519	0.4224	0.9612
A. graveolens-*V. vinifera* (speciation)	1.3022	0.1878	0.9686
C. asiatica-*V. vinifera* (gamma)	1.5838	0.2906	0.9431
C. asiatica-*V. vinifera* (speciation)	1.1554	0.1357	0.9766
E. senticosus-*V. vinifera* (gamma)	1.4135	0.2386	0.9719
E. senticosus-*V. vinifera* (speciation)	0.9966	0.1098	0.983
P. notoginseng-*V. vinifera* (gamma)	1.3906	0.2103	0.9516
P. notoginseng-*V. vinifera* (speciation)	1.0036	0.0868	0.9433
P. vietnamensis var. *fuscidiscus*-*V. vinifera* (gamma)	1.4178	0.1673	0.9193
P. vietnamensis var. *fuscidiscus*-*V. vinifera* (speciation)	1.0055	0.1114	0.9448
A. graveolens-*C. asiatica* (gamma)	1.8155	0.2772	0.9523

A. graveolens-*C. asiatica* (speciation)	0.8394	0.1392	0.9548
P. vietnamensis var. *fuscidiscus*-*A. graveolens* (gamma)	1.7051	0.2832	0.9618
P. vietnamensis var. *fuscidicu* -*A. graveolens* (speciation)	0.7184	0.0922	0.9781
P. vietnamensis var. *fuscidiscus*-*C. asiatica* (gamma)	1.6343	0.248	0.9469
P. vietnamensis var. *fuscidiscus*-*C. asiatica* (speciation)	0.5348	0.0734	0.9618
P. vietnamensis var. *fuscidiscus*-*E. senticosus* (gamma)	1.4891	0.1774	0.9418
P. vietnamensis var. *fuscidicu* -*E. senticosus* (Pg-β)	0.3746	0.0351	0.9637
P. vietnamensis var. *fuscidiscus*-*E. senticosus* (speciation)	0.1476	0.0222	0.9772
P. vietnamensis var. *fuscidiscus*-*P. notoginseng* (gamma)	1.4978	0.2106	0.9145
P. vietnamensis var. *fuscidiscus*-*P. notoginseng* (Pg-β)	0.3801	0.0404	0.9577
P. vietnamensis var. *fuscidiscus*-*P. notoginseng* (speciation)	0.0311	0.0094	0.9839

Table S18. Summary of Collinear genomics subsets for five species.

Species	***V. vinifera***	***C. asiatica***	***E. senticosus***	***P. notoginseng***	***P. vietnamensis* var. *fuscidiscus***
Whole genome duplications	γ triplication	γ triplication Casi-α/Apiaceae-common WGD	γ triplication Pg-β Esen-α	γ triplication Pg-β -	γ triplication Pg-β -
Number of collinear copies	3	6	12	6	6

Table S19. General statistics on oxidosqualene cyclase genes from eight species identified using HMMER.

Species	Sequence id	Name	CDS length (bp)	Function/Putative function
A. trichopoda	ERN12565	AtOSC1	2,286	Cycloartenol synthase
A. fimbriata	KAG9449364.1	AfOSC1	2,283	Cycloartenol synthase
	KAG9449355.1	AfOSC2	2,265	Lanosterol synthase
V. vinifera	GSVIVT01029468001	VvOSC1	1,683	Cycloartenol synthase
	GSVIVT01029527001	VvOSC2	2,274	Cycloartenol synthase
	GSVIVT01029514001	VvOSC3	2,274	Cycloartenol synthase
	GSVIVT01029525001	VvOSC4	2,409	Cycloartenol synthase
	GSVIVT01029524001	VvOSC5	2,409	Cycloartenol synthase
	GSVIVT01015994001	VvOSC6	1,683	Cycloartenol synthase
	GSVIVT01032285001	VvOSC7	2,283	Cycloartenol synthase
	GSVIVT01032217001	VvOSC8	2,406	Lupeol synthase
	GSVIVT01021473001	VvOSC9	2,262	β-amyrin synthase and other mTTSs (Group I)
	GSVIVT01021474001	VvOSC10	2,280	β-amyrin synthase and other mTTSs (Group I)
	GSVIVT01021494001	VvOSC11	2,280	β-amyrin synthase and other mTTSs (Group I)
	GSVIVT01021495001		24,726	β-amyrin synthase and other mTTSs (Group I)
	GSVIVT01029510001	VvOSC12	2,445	β-amyrin synthase and other mTTSs (Group II)
	GSVIVT01029509001	VvOSC13	2,541	β-amyrin synthase and other mTTSs (Group II)
	GSVIVT01029491001	VvOSC14	1,539	β-amyrin synthase and other mTTSs (Group II)
	GSVIVT01029488001	VvOSC15	2,310	β-amyrin synthase and other mTTSs (Group II)
	GSVIVT01029489001	VvOSC16	2,619	β-amyrin synthase and other mTTSs (Group II)
	GSVIVT01029508001	VvOSC17	2,733	β-amyrin synthase and other mTTSs (Group II)
	GSVIVT01029474001	VvOSC18	2,658	β-amyrin synthase and other mTTSs (Group II)
C. asiatica	evm.model.Scaffold_7.3187	CaOSC1	2,274	Cycloartenol synthase
	evm.model.Scaffold_3.13	CaOSC2	2,391	Lanosterol synthase

	evm.model.Scaffold_1.5050	CaOSC3	4,389	Lupeol synthase
	evm.model.Scaffold_1.163	CaOSC4	2,283	β-amyrin synthase and other mTTSs (Group I)
	evm.model.Scaffold_1.3735	CaOSC5	2,283	β-amyrin synthase and other mTTSs (Group I)
	evm.model.Scaffold_1.3299	CaOSC6	2,298	β-amyrin synthase and other mTTSs (Group I)
	evm.model.Scaffold_7.2291	CaOSC7	2,286	β-amyrin synthase and other mTTSs (Group II)
	evm.model.Scaffold_5.2605	CaOSC8	2,298	β-amyrin synthase and other mTTSs (Group II)
	evm.model.Scaffold_5.2606	CaOSC9	2,295	β-amyrin synthase and other mTTSs (Group II)
	evm.model.Scaffold_7.2982	CaOSC10	2,178	β-amyrin synthase and other mTTSs (Group II)
E. senticosus	Ese03G002792.t1	EsOSC1	2,277	Cycloartenol synthase
	Ese12G002640.t1	EsOSC2	2,274	Cycloartenol synthase
	Ese18G000732.t1	EsOSC3	2,487	Lupeol synthase
	Ese07G000079.t1	EsOSC4	2,286	β-amyrin synthase and other mTTSs (Group II)
	Ese17G001441.t1	EsOSC5	2,160	β-amyrin synthase and other mTTSs (Group II)
	Ese24G002150.t1	EsOSC6	2,253	β-amyrin synthase and other mTTSs (Group II)
	Ese11G000382.t1	EsOSC7	2,295	β-amyrin synthase and other mTTSs (Group II)
	Ese11G000379.t1	EsOSC8	1,752	β-amyrin synthase and other mTTSs (Group II)
P. notoginseng	Pno05G000040.t1	PnOSC1	2,274	Cycloartenol synthase
	Pno01G006888.t1	PnOSC2	2,277	Cycloartenol synthase
	Pno05G000039.t1	PnOSC3	2,331	Lanosterol synthase
	Pno05G003783.t1	PnOSC4	2,178	Lupeol synthase
	Pno10G001026.t1	PnOSC5	2,280	β-amyrin synthase and other mTTSs (Group I)
	Pno03G005730.t1	PnOSC6	2,310	Dammarenediol synthase
	Pno03G005732.t1	PnOSC7	2,310	Dammarenediol synthase
	Pno02G006287.t1	PnOSC8	2,286	β-amyrin synthase and other mTTSs (Group II)
	Pno04G000589.t1	PnOSC9	2,274	β-amyrin synthase and other mTTSs (Group II)
	Pno09G004398.t1	PnOSC10	2,316	β-amyrin synthase and other mTTSs (Group II)
	Pno12G004330.t1	PnOSC11	2,292	β-amyrin synthase and other mTTSs (Group II)
	Pno02G002482.t1	PnOSC12	1,977	β-amyrin synthase and other mTTSs (Group II)
P. vietnamensis	Pvi05G002838.t1	PvOSC1	2,274	Cycloartenol synthase

	Pvi79G000004.t1	PvOSC2	1,977	Cycloartenol synthase
	Pvi04G017143.t1	PvOSC3	2,277	Cycloartenol synthase
	Pvi05G002839.t1	PvOSC4	2,334	Lanosterol synthase
	Pvi05G007905.t1	PvOSC5	2,187	Lupeol synthase
	Pvi08G000242.t1	PvOSC6	2,280	β-amyrin synthase and other mTTSs (Group I)
	Pvi06G002447.t1	PvOSC7	2,310	Dammarenediol synthase
	Pvi02G002534.t1	PvOSC8	2,286	β-amyrin synthase and other mTTSs (Group II)
	Pvi07G000589.t1	PvOSC9	2,289	β-amyrin synthase and other mTTSs (Group II)
	Pvi12G004397.t1	PvOSC10	2,361	β-amyrin synthase and other mTTSs (Group II)
	Pvi01G001914.t1	PvOSC11	2,337	β-amyrin synthase and other mTTSs (Group II)
P. ginseng	Pg_S0266.37	PgOSC1	2,277	Cycloartenol synthase
	Pg_S0762.36	PgOSC2	2,277	Cycloartenol synthase
	Pg_S0910.3	PgOSC3	2,490	Cycloartenol synthase
	Pg_S2798.13	PgOSC4	2,277	Cycloartenol synthase
	Pg_S0701.10	PgOSC5	2,205	Cycloartenol synthase
	Pg_S0266.35	PgOSC6	2,214	Lanosterol synthase
	Pg_S0762.35	PgOSC7	2,331	Lanosterol synthase
	Pg_S0577.13	PgOSC8	2,280	Lupeol synthase
	Pg_S4166.7	PgOSC9	2,289	β-amyrin synthase and other mTTSs (Group I)
	Pg_S3517.9	PgOSC10	1,524	Dammarenediol synthase
	Pg_S3318.3	PgOSC11	2,310	Dammarenediol synthase
	Pg_S4815.4	PgOSC12	2,286	β-amyrin synthase and other mTTSs (Group II)
	Pg_S0034.9	PgOSC13	2,286	β-amyrin synthase and other mTTSs (Group II)
	Pg_S0034.2	PgOSC14	2,295	β-amyrin synthase and other mTTSs (Group II)
	Pg_S2801.2	PgOSC15	2,286	β-amyrin synthase and other mTTSs (Group II)
	Pg_S2939.4	PgOSC16	2,289	β-amyrin synthase and other mTTSs (Group II)
	Pg_S0361.30	PgOSC17	2,043	β-amyrin synthase and other mTTSs (Group II)
	Pg_S2492.7	PgOSC18	2,292	β-amyrin synthase and other mTTSs (Group II)
	Pg_S0888.6	PgOSC19	2,292	β-amyrin synthase and other mTTSs (Group II)

	AB009029	PgPNX1	2,259	Cycloartenol synthase
	AB009030	PgPNY1	2,589	β-amyrin synthase
	AB009031	PgPNZ1	2,684	Lanosterol synthase
Welwitschia mirabilis	W.mirabilis.02578	WmOSC1	2,280	Cycloartenol synthase
Artemisia annua	KM670094	AaLUS	2,274	Lupeol synthase
Panax quinquefolius	GU997679	PqDDS	2,310	Dammarenediol synthase

Table S20. Product profile for OSCs based on GC analysis. √ and × represents presence and absence of compounds.

Protein ID	1	2	3	5	6	7	8	9	*	#	Unidentified
PvOSC6	√	√	√	√	√	√	√	√	√	√	√
PgOSC9	√	√	√	√	√	√	√	√	√	√	√
PnOSC5	√	√	√	√	√	√	√	√	√	√	√
CaOSC6	√	√	√	√	√	√	√	√	√	√	√
CaOSC5	√	√	√	√	√	√	√	√	√	√	√
PgOSC11	×	×	×	×	×	√	√	√	√	√	×
PqDDS	×	×	×	×	×	√	√	√	√	√	×
PvOSC7	×	×	×	×	×	√	√	√	√	√	×
PnOSC6	×	×	×	×	×	√	√	√	√	√	×

Table S21. ^{13}C and ^{1}H NMR assignments for compound 8 and 9.

No.	**Compound 8** (3-epicabraleadiol)		**Compound 9** (ocotillol)	
	δH (J in Hz)	δC	δH (J in Hz)	δC
C1(CH_2)	1.69(1H, m), 0.97(1H, m)	39.04	1.67(1H, m), 0.95(1H, m)	39.03
C2(CH_2)	1.65(1H, m), 1.57(1H, m)	27.41	1.64(1H, m), 1.57(1H, m)	27.39
C3(CH)	3.20 (1H, dd, *J* = 11.5, 4.7)	78.95	3.19 (1H, dd, *J* = 11.5, 4.8)	78.93
C4(qC)		38.96		38.95
C5(CH)	0.73 (1H, dd, J = 11.9, 2.2)	55.84	0.72 (dd, J = 12.0, 2.0)	55.83
C6(CH_2)	1.52 (1H, m), 1.421H, (m)	18.27	1.50(1H, m), 1.44(1H, m)	18.25
C7(CH_2)	1.53(1H, m), 1.27(1H, m)	35.27	1.51(1H, m), 1.26(1H, m)	35.25
C8(qC)		40.37		40.34
C9(CH)	1.32 (1H, m)	50.81	1.31(1H, m)	50.76
C10(qC)		37.14		37.12
C11(CH_2)	1.87(1H, m), 1.50(1H, m)	21.8	1.49(1H, m),1.46(1H, m)	21.53
C12(CH_2)	1.75(1H, m), 1.33(1H, m)	25.85	177(1H, m),1.47(1H, m)	25.7
C13(CH)	1.64 (1H, m)	42.83	1.56 (1H, m)	42.93
C14(qC)		50.02		50.03
C15(CH_2)	1.46 (1H, m), 1.06 (1H, m)	26.99	1.62(1H, m), 1.44(1H, m)	26.1
C16(CH_2)	1.76 (1H, m), 1.63 (1H, m)	31.44	1.84 (1H, m), 1.06 (1H, m)	31.44
C17(CH)	1.86 (1H, m)	49.82	1.79 (1H, m)	49.5
C18(CH_3)	0.97(3H, s)	15.47	0.94(3H, s)	15.41
C19(CH_3)	0.85(3H, s)	16.24	0.83(3H, s)	16.21
C20(qC)		86.55		86.41
C21(CH_3)	1.11(3H, s)	27.18	1.12(3H, s)	23.51
C22(CH_2)	1.87(1H, m), 1.67(1H, m)	34.72	1.62(1H, m), 1.55(1H, m)	35.66
C23(CH_2)	1.80(1H, m), 1.75(1H, m)	26.35	1.85(1H, m), 1.77(1H, m)	27.35
C24(CH)	3.64 (dd, *J* = 10.2, 5.2)	86.29	3.72 (t, J = 7.4)	83.29
C25(qC)		70.24		71.43
C26(CH_3)	1.19(3H, s)	27.83	1.20(3H, s)	27.46

C27(CH_3)	1.11(3H, s)	24.04	1.11(3H, s)	24.23
C28(CH_3)	0.97 (3H, s)	27.99	0.96 (3H, s)	27.97
C29(CH_3)	0.77(3H, s)	15.35	0.76(3H, s)	15.33
C30(CH_3)	0.87(3H, s)	16.4	0.86(3H, s)	16.44

Table S22. Relative composition of identified products for the nine OSCs. The relative abundance of each compound is calculated based on the area of the corresponding peak.

Protein ID	δ-Amyrin	β-Amyrin	α-Amyrin	ψ-Taraxasterol	Taraxasterol	Dammarenediol-II	3-Epicabraleadiol	Ocotillol	Total
PvOSC6	1.35%	27.41%	52.91%	3.35%	0.81%	0.19%	12.99%	0.99%	100.00%
PgOSC9	0.93%	24.29%	63.06%	3.57%	0.46%	0.52%	5.86%	1.31%	100.00%
PnOSC5	1.01%	23.71%	62.46%	3.12%	0.71%	0.16%	7.90%	0.93%	100.00%
CaOSC6	3.34%	23.78%	8.51%	43.33%	7.63%	0.88%	11.37%	1.16%	100.00%
CaOSC5	1.20%	29.97%	60.47%	3.48%	0.51%	0.10%	3.82%	0.45%	100.00%
PgOSC11	0%	0%	0%	0%	0%	22.82%	4.77%	72.41%	100.00%
PqDDS	0%	0%	0%	0%	0%	80.25%	2.04%	17.71%	100.00%
PvOSC7	0%	0%	0%	0%	0%	80.05%	2.05%	17.90%	100.00%
PnOSC6	0%	0%	0%	0%	0%	84.11%	1.85%	14.04%	100.00%

Table S23. Summary of *P. vietnamensis* var. *fuscidiscus* sequencing data.

Type	Library	Insert size (bp)	Reads number	GC content (%)	Mean reads length (bp)	Reads length N50 (bp)	Base (Gb)	Coverage depth
Illumina genomic sequencing	YSQ-1	150	207,536,126	39.98	-	-	31.13	132.64 X
	YSQ-3	150	421,194,574	40.01	-	-	63.18	
	YSQ-5	150	370,979,332	40.03	-	-	55.65	
	YSQ-7	150	530,083,054	39.14	-	-	79.51	
Pacbio genomic sequencing	YSQ-pacbio	-	11,577,317	-	10,117	17,312	117.13	67.71 X
Hic sequencing	YSQ-Hic	150	1,688,961,966	37	-	-	253.34	146.44 X
RNA sequencing	Leaf-1	150	49,842,528	-	150	150	7.46	12.32 X
	Stem-1	150	43,288,694	-	150	150	6.49	
	Root-1	150	49,166,928	-	150	150	7.36	
Total	-	-	-	-	-	-	621.25	370.94 X

Table S24. Source information of species used in phylogenetic analysis.

Species	Family	Source
Vitis vinifera	Vitaceae	10.1038/nature06148
Coffea canephora	Rubiaceae	10.1126/science.1255274
Codonopsis pilosula	Campanulaceae	medicinalplants.ynau.edu.cn/genome
Welwitschia mirabilis	Welwitschiaceae	10.1038/s41467-021-24528-4
Lactuca sativa	Asteraceae	10.1038/ncomms14953
Lonicera japonica	Lonicera japonica	10.1111/nph.16552
Centella asiatica	Apiaceae	10.1016/j.ygeno.2021.05.019
Daucus carota	Apiaceae	10.1038/ng.3574
Apium graveolens	Apiaceae	10.1111/pbi.13499
Eleutherococcus senticosus	Araliaceae	10.1111/1755-0998.13403
Panax ginseng	Araliaceae	10.1111/pbi.12926
Panax notoginseng	Araliaceae	This study
Panax vietnamensis var. *fuscidiscus*	Araliaceae	This study