Plant Biotechnology Journal

WILEY

RESEARCH ARTICLE OPEN ACCESS

Chromosome-Level Genome Assembly of the Allotetraploid *Gynostemma pentaphyllum* Provides Novel Insights Into the Biosynthesis of Ginsenoside and Gypenoside LVI

Peina Zhou[1,2] | Si-Jie Liu[3] | Lijin Huang[1] | Yingping Wang[1] | Xinyu Jiang[4,5] | Wang Dong[3] | Jianfeng Gong[5] | Long Wang[1] | Yuyin Zhao[1] | Huiying Wang[1] | Ping Li[1] | Jia-Yu Xue[3] | Xu Lu[1,6]

[1]State Key Laboratory of Natural Medicines, School of Traditional Chinese Pharmacy, China Pharmaceutical University, Nanjing, China | [2]Nanjing Research Institute for Comprehensive Utilization of Wild Plants, Nanjing, China | [3]College of Horticulture, Bioinformatics Center, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing, China | [4]Hunan Huabaotong Pharmaceutical Co. LTD, Changsha, China | [5]Wepon Pharmaceutical Holding Group Co. Ltd, Hangzhou, China | [6]State Key Laboratory for Quality Ensurance and Sustainable Use of Dao-di Herbs, Beijing, China

Correspondence: Ping Li (liping2004@126.com) | Jia-Yu Xue (xuejy@njau.edu.cn) | Xu Lu (luxu666@163.com)

Received: 14 October 2025 | **Revised:** 7 February 2026 | **Accepted:** 11 February 2026

Keywords: convergent evolution | cytochrome P450 | ginsenoside | *Gynostemma pentaphyllum* | gypenoside LVI | haploid-resolved allotetraploid chromosome-level genome

ABSTRACT

Gynostemma pentaphyllum, a herb used in tea and traditional Chinese medicine, shows geographic variation in its production of valuable dammarane-type ginsenosides and gypenoside LVI between populations from Suining (SN) and Nanning (NN). To elucidate the mechanisms underlying this differential metabolite accumulation, a chromosome-level genome for *G. pentaphyllum* (SN population) was assembled. The analysis revealed that SN is a tetraploid (~1.2 Gb), resulting from a recent whole-genome duplication event in a diploid ancestor. Phylogenetic analysis indicates SN and diploid NN share a recent common ancestor, diverging approximately 4.95 million years ago. Chromosome evolution analysis confirmed SN is an allotetraploid with clear subgenomic differentiation. This genome, combined with multi-omics data, enabled the screening of candidate P450 genes involved in ginsenoside/gypenoside LVI biosynthesis. In vivo and in vitro experiments confirmed that GpCYP88AB3 functions as a bifunctional enzyme by first hydroxylating dammarenediol-II at C-12 to yield protopanaxadiol (PPD), and then hydroxylating PPD at C-2 to form 2α-OH-PPD. Phylogenetically, GpCYP88AB3 and similar enzymes from Araliaceae belong to distinct CYP subfamilies, demonstrating convergent evolution of this function between the two plant families and highlighting the functional plasticity of P450s. Evolutionary analysis suggests that GpCYP88AB3 emerged from a CYP88 gene family expansion in the tetraploid *G. pentaphyllum*. This expansion occurred after, but was not directly caused by, the whole-genome duplication event. This study elucidates the biosynthetic pathway for the key metabolites in *G. pentaphyllum*, providing a foundation for future metabolic engineering and synthetic biology applications.

1 | Introduction

Gynostemma pentaphyllum (Thunb.) Makino, a perennial herbaceous vine within the Cucurbitaceae family, is naturally distributed across Asia. Its native range includes China, Japan, South Korea, and Vietnam, with naturalised populations extending into southern and southeastern regions such as Bangladesh, Malaysia and Indonesia (Zhang et al. 2023). Within China, wild

Peina Zhou, Si-Jie Liu, Lijin Huang and Yingping Wang contributed equally to this work.

populations are particularly abundant throughout the Yangtze River Basin and southern provinces, including Fujian, Guangxi, Hunan and Jiangsu. Extensive pharmacological studies have demonstrated that *G. pentaphyllum* possesses diverse therapeutic properties, including antimicrobial, anticancer, anti-aging, anti-fatigue, anti-ulcer, lipid-lowering and immunomodulatory activities (Gou et al. 2018; Lee et al. 2019; Megalli et al. 2006; Zu et al. 2021). These multifaceted effects are primarily attributed to its rich content of dammarane-type ginsenosides and gypenosides. Owing to its remarkable therapeutic potential, *G. pentaphyllum* is not only a traditional medicinal herb but also a key ingredient in herbal teas and functional foods, underscoring its considerable pharmaceutical and commercial value (Chen, Guo, et al. 2023). Under natural growing conditions, *G. pentaphyllum* exhibits a complex multiploid system comprising diploid, tetraploid, hexaploid and octoploid cytotypes. Notably, elevated ploidy levels are associated with increased plant biomass and amplified biosynthesis of secondary metabolites across various tissues (An et al. 2024). Polyploidisation, a key evolutionary mechanism in plants, drives genomic novelty through whole-genome duplication (WGD). This process facilitates functional diversification of alleles, emergence of novel metabolic pathways, and phenotypic innovation, underscoring its critical role in the evolution of *G. pentaphyllum* (Paterson and Queitsch 2024).

Among the diverse phytochemical constituents of *G. pentaphyllum*, gypenosides are the principal bioactive agents responsible for its multifunctional pharmacological properties. These dammarane-type tetracyclic triterpenoid saponins represent a structurally unique class of natural products, with over 300 distinct variants isolated and characterised from this species to date (Nguyen et al. 2021; Su et al. 2021). Notably, approximately 25% of these compounds exhibit structural homology with protopanaxadiol (PPD)-type ginsenosides, with specific examples including identity between Gypenoside III and Ginsenoside Rb1, Gypenoside IV and Ginsenoside Rb3, Gypenoside VIII and Ginsenoside Rd, as well as Gypenoside XII and Ginsenoside F2 (Nguyen et al. 2021; Su et al. 2021). Consequently, *G. pentaphyllum* is colloquially termed 'Southern Ginseng' in China due to this phytochemical equivalence and has emerged as a strategic botanical alternative for bioactive ginsenoside production. Additionally, *G. pentaphyllum* synthesises characteristic 2α-OH-PPD type gypenosides (e.g., gypenoside XLVI, gypenoside LVI and gypenoside L), which are considered signature components of the *Gynostemma* genus (Huang et al. 2022; Lou et al. 2021; Zhang et al. 2015). The accumulation of gypenosides exhibits significant geographic variation. For instance, chemoprofiles diverge considerably among accessions from Guangxi, Fujian, Hunan, Sichuan and Yunnan, resulting in distinct flavour profiles and pharmacological activities (Chen, Yao, et al. 2023; Liang et al. 2025; Wang et al. 2020).

Dammarane-type tetracyclic triterpenoid saponins are biosynthesised via the mevalonic acid (MVA) and methylerythritol 4-phosphate (MEP) pathways. Both pathways converge to synthesise the universal five-carbon isoprenoid precursors, isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP). Geranyl diphosphate synthase (GPPS) catalyses their condensation to form geranyl diphosphate (GPP), which farnesyl diphosphate synthase (FPPS) extends to farnesyl diphosphate (FPP). Sequential action of squalene synthase (SS) and squalene epoxidase (SE) converts FPP into 2,3-oxidosqualene (Liu et al. 2025; Yun et al. 2024). Dammarenediol-II synthase (DS) then cyclises this intermediate to dammarenediol-II. Then, cytochrome P450 enzymes (CYP450s) might subsequently catalyse C-12 hydroxylation of dammarenediol-II to yield protopanaxadiol (PPD) and C-2 hydroxylation to yield 2α-OH-PPD. UDP-glycosyltransferases (UGTs) elaborate these aglycones into diverse triterpenoid saponins through site-specific glycosylation using sugar donors (Yun et al. 2024). Although genes encoding intermediate pathway enzymes (e.g., GpDS, GpUGT71AW1, GpUGT74AC3, GpUGT71V3) have been functionally validated, the specific CYP450 genes catalysing the formation of the critical precursors PPD and 2α-OH-PPD remain unidentified in *G. pentaphyllum* (Le et al. 2021; Yun et al. 2024). Systematic profiling of CYP450 genes using a high-quality, chromosome-level genome assembly is therefore pivotal for elucidating gypenoside biosynthesis in this species. Characterising this pathway at the gene level will establish a foundation for metabolic engineering strategies to enhance the targeted production of these pharmacologically active saponins (Li et al. 2024).

Here, we collected *G. pentaphyllum* populations of Suining (SN population, tetraploid) and Nanning (NN population, diploid) that were detected to contain different content of gypenosides, and conducted comprehensive analyses, including comparative genomics, transcriptomics and metabolomes, to explore the mechanisms underlying the phenomenon. Through multi-omics screening, functional characterisation, and evolutionary analyses, we identified GpCYP88AB3 as uniquely catalysing the sequential C-12 and C-2 hydroxylation of dammarenediol-II to form protopanaxadiol (PPD) and 2α-OH-PPD. We further revealed its independent origin via convergent evolution relative to Araliaceae ginseng P450s. This integrated investigation establishes a framework for reconstructing the complete gypenoside biosynthetic pathway and provides a scalable synthetic biology chassis for combinatorial optimisation of triterpenoid saponin production.

2 | Results

2.1 | Distinct Metabolomic Profiles Between SN and NN *G. pentaphyllumn* Populations

To evaluate ginsenoside and 2α-OH-PPD type gypenosides accumulation across *G. pentaphyllum* populations, we analysed roots, stems, and leaves from SN and NN populations using LC-QTOF-MS/MS. Total ion chromatograms and principal component analysis (PCA) revealed significant chemotypic divergence between populations and tissues (Figure S1). By matching precursor and fragment ions with authentic standards (Rb1, Rd, Rb3, Compound K, F2, Rg3, Rg5, Gypenoside L, XLVI, and LVI), we quantified all 10 target compounds in tissues of SN and NN (Figure S1, Table S1). PPD-type ginsenosides (Rb1, Rb3, Rd, F2, Rg3, Rg5 and Compound K) and 2α-OH-PPD-type gypenosides (L, XLVI, and LVI) accumulated predominantly in tissues of SN. Gypenosides LVI and L were exclusive to SN, while XLVI occurred in tissues of SN and roots of NN. Quantitative profiling of 10 gypenoside in *G. pentaphyllum* from SN and NN revealed distinct accumulation patterns. Specifically, gypenoside LVI was the most abundant in SN, with mean concentrations of $1114.53 \pm 311.47\,\mu g/g$

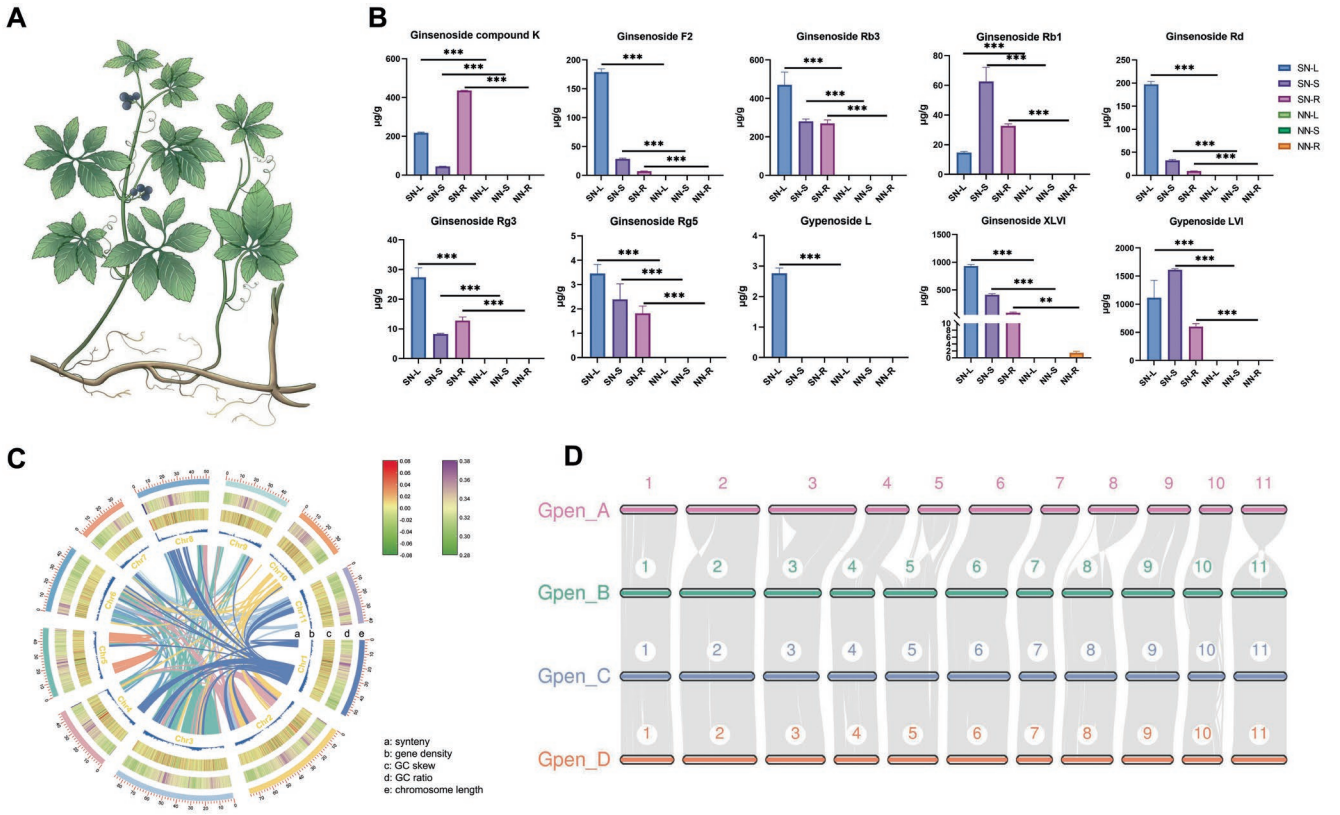

FIGURE 1 | The contents of main gypenosides and genome assembly analysis of SN *G. pentaphyllum*. (A) Morphology of SN *G. pentaphyllum*. (B) The contents of main gypenosides in different tissues of SN and NN *G. pentaphyllum*. (C) Genomic features of SN *G. pentaphyllum*. The circus plot from the outer to the inner circle represents chromosome-scale pesudochromosomes (Chr1-Chr11) (a), gene density (b), the density of repeat sequences (c), GC content (d), and each linking line in the center of the circus plot indicates a pair of homologous genes (e). (D) Synteny analysis of the genomes of four haplotype of SN *G. pentaphyllum*. The links represent one-to-one synteny homologues between the species. The *** represents the $p < 0.001$, and ** represents the $p < 0.01$.

in leaves, 1612.58 ± 17.56 mg/g in stems, and 602.18 ± 54.22 μg/g in roots. Gypenoside XLVI ranked second, showing mean concentrations of 930.85 ± 27.77 μg/g in leaves, 412.59 ± 20.19 μg/g in stems, and 88.93 ± 12.68 μg/g in roots, while its level in roots of NN was significantly lower (2.22 ± 0.63 μg/g). Gypenoside L showed highest accumulation in leaves of SN (25.67 ± 1.86 μg/g), with lower levels in stems and roots. Rb3 was the most abundant ginsenoside in SN, exhibiting a clear tissue gradient: leaves > stems > roots. Rd accumulated predominantly in leaves, with minimal levels in stems and roots (Figure 1A,B). Detailed concentrations of F2, Rb1, Rg3, CK and Rg5 are provided in Table S2.

2.2 | Genome Assembly and Annotation of the SN Tetroploid *G. pentaphyllum*

According to the previous report, NN is a diploid *G. pentaphyllum* (Huang et al. 2022). Given the significant chemotypic divergence in PPD-type ginsenosides and 2α-OH-PPD-type gypenosides between SN and NN, we performed karyotype analysis and flow cytometry of SN. The results indicated that SN is a tetraploid *G. pentaphyllum*, exhibited 44 chromosomes ($2n = 4\times = 44$) and a ~1.2-Gb genome size (Table S3; Figure S2). Leveraging deep-coverage sequencing (107 Gb PacBio HiFi reads; 100× Hi-C data), we generated a comprehensive haplotype-phased genome assembly. Initial hifiasm assembly (–hic mode) yielded four high-contiguity haplotypes: Gpen_A (580.62 Mb; 45 contigs; N50 = 65.88 Mb), Gpen_B (618.42 Mb; 66 contigs; N50 = 71.97 Mb), Gpen_C (643.42 Mb; 150 contigs; N50 = 31.91 Mb), and Gpen_D (648.41 Mb; 154 contigs; N50 = 28.02 Mb) (Table S4). Hi-C scaffolding anchored each haplotype to 11 pseudochromosomes with >98% anchoring rates (Figures S2 and S3). Integrated gene annotation (ab initio prediction, homology-based methods, RNA-seq evidence) identified 33 559 (Gpen_A), 36 943 (Gpen_B), 39 170 (Gpen_C), and 38 577 (Gpen_D) protein-coding genes. Repetitive sequences constituted 64.51%–63.65% of assemblies, predominantly LTR retrotransposons (23.18%–24.83%) (Figure 1C). BUSCO assessments (embryophyta_odb10) confirmed assembly completeness: 98.4% (Gpen_A), 98.6% (Gpen_B), 99.0% (Gpen_C), 98.9% (Gpen_D); and annotation completeness: 93.1% (Gpen_A), 92.2% (Gpen_B), 92.8% (Gpen_C) and 92.8% (Gpen_D) (Table S5). This haplotype-resolved genome provides unprecedented resolution for exploring polyploid evolution and dammarane-type saponin pathway diversification in *G. pentaphyllum*.

2.3 | Evolutionary Analyses of Tetraploid SN *G. pentaphyllum*

To elucidate the phylogenetic position of tetraploid *G. pentaphyllum*, a time-calibrated phylogeny was reconstructed

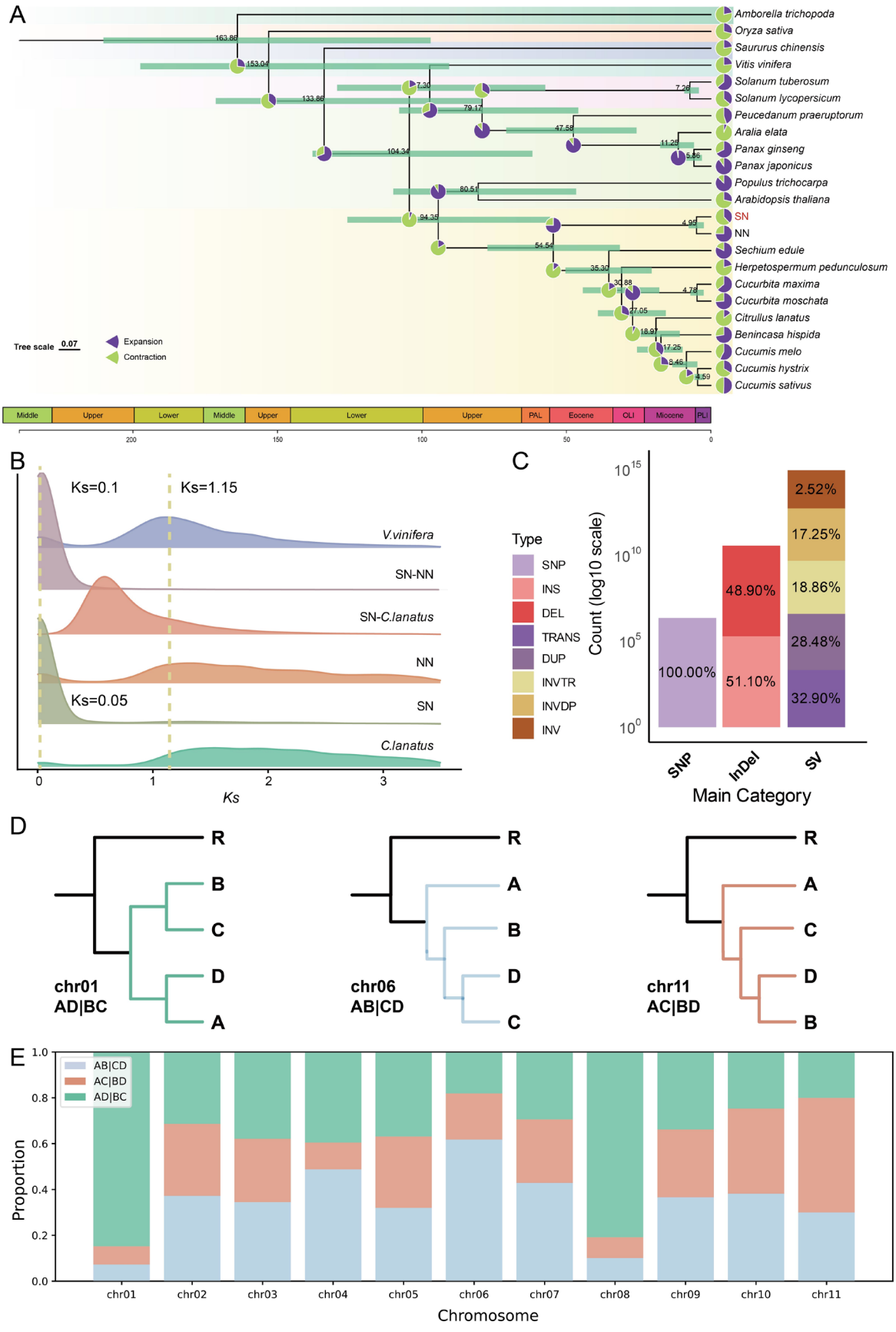

FIGURE 2 | Legend on next page.

FIGURE 2 | Evolutionary analyses of SN *G. pentaphyllum* (tetraploid). (A) The species tree displays the divergence times between tetraploid *G. pentaphyllum* and the other species. The gene families that have expanded and contracted are depicted in purple and green, respectively. (B) WGD analysis of SN *G. pentaphyllum* (tetraploid). (C) Genomic structural variations of SN (tetraploid) and NN (diploid) *G. pentaphyllum*. (D) Three alternative quartet topologies (AB|CD, AC|BD, AD|BC) that partition the four subgenomes into two pairs, with the diploid reference (R) as outgroup. Unlike standard phylogenetic tree topologies that resolve all taxa simultaneously, quartet-based topologies only compare relationships among four taxa at a time, thus providing localised and statistically testable signals of subgenome affinity. (E) Stacked bar plots summarising quartet topology frequencies across chromosomes. Each bar represents one chromosome, partitioned into AB|CD, AC|BD, AD|BC categories. This highlights chromosome-specific variation in phylogenetic signal, such as dominant AD|BC topology on chr01–chr03, AB|CD on chr06–chr07, and mixed or weak signals on chr08–chr11.

using 23 representative angiosperms (Figure 2A; Table S6). The analysis revealed that *G. pentaphyllum* represents the earliest-diverging lineage among the used Cucurbitaceae species, with an estimated divergence time of approximately 54.54 million years ago (Mya). In addition, the phylogenetic tree shows that tetroploid SN and diploid NN share a recent common ancestor, and the divergence time between the tetroploid and diploid *G. pentaphyllum* is estimated to be around 4.95 Mya. Furthermore, the phylogenetic tree indicates that *G. pentaphyllum* is phylogenetically distant from Araliaceae species, such as *Aralia elata*, *Panax ginseng* and *P. japonicus*, and their most recent common ancestor should date back to at least 104.34 Mya. Gene family dynamics analysis identified 10 681 expanded and 385 contracted gene families in the tetraploid SN, exhibiting a significant expansion in gene families compared to its diploid counterpart. Such lineage-specific genomic innovations should be associated with its polyploid nature, and possibly contributed to the metabolic diversity and adaptation of the tetraploid SN.

To investigate whole-genome duplication (WGD) events in the tetraploid SN, we analysed the synonymous substitution rate (Ks) of homologous gene pairs. The Ks distribution revealed two distinct peaks (Figure 2B). The first peak (Ks≈0.05) corresponded to the most recent species-specific tetraploidisation event, and the second peak (Ks≈1.15) was in line with the γ-whole-genome triplication event shared by core eudicots. The additional peak at Ks≈0.1 reflects divergence between tetraploid SN and diploid NN, suggesting that tetraploid SN formed via an independent WGD event in the SN lineage after its divergence from the NN lineage.

Macro-scale collinearity analysis revealed pronounced genomic conservation between the diploid genome (Huang et al. 2021; Yun et al. 2024) and the newly assembled tetraploid SN genome (Figure S4), with most genic regions retaining high syntenic conservation. Nevertheless, discrete structural rearrangements were detected on specific chromosomes: Chromosome 1 exhibited small-scale segmental duplications (blue arcs), suggesting gene family expansion or tandem repeat proliferation. Chromosome 5 harboured distinct inversions (orange arcs) and translocations (green arcs), potentially indicative of species- or subspecies-level genomic divergence within these regions. These findings underscore the dynamic interplay between conserved synteny and lineage-specific structural reorganisation in shaping the *G. pentaphyllum* genome architecture. Extending beyond structural conservation patterns, our comparative genomic analysis between the newly assembled chromosome-level tetraploid SN genome and the published Telomere-to-Telomere (T2T) reference genome of *G. pentaphyllum* identified distinct genetic variation profiles (Figure S4A). Single nucleotide polymorphisms (SNPs) constituted the predominant class of genomic variation. Within the InDel category, insertions (INS) and deletions (DEL) demonstrated comparable prevalence (51.10% vs. 48.90%, respectively), reflecting equilibrium between insertion and deletion events. Structural variation (SV) analysis further established translocations (TRANS, 32.90%) and duplications (DUP, 28.48%) as the dominant SV types (Figure 2C), which may influence gene expression through dosage effects or alterations in regulatory landscapes.

In the phylogenetic inference, different chromosomes exhibited distinct topologies. The phylogenetic tree of chr01 showed the topology (R,((B,C),(D,A))), indicating clustering of B with C and A with D; chr03 and chr05 exhibited the topology (R,((D,(B,C)),A)), further supporting the close relationship between B and C; chr06 displayed (R,(A,(B,(D,C)))) (Figure 2D, Table S7a), highlighting the affinity between C and D. In contrast, chr11 contained fewer effective gene trees, resulting in a weaker phylogenetic signal. Quartet statistics and visualisation further validated these patterns. Chromosomes chr01–chr03 showed significantly higher support for the AD|BC topology, chr06 and chr07 better fit the AB|CD topology, while chr08 and chr09 presented mixed signals, suggesting the possibility of complex rearrangements or gene flow (Figure 2E). The statistical results for chr11 were weak (Table S7b), indicating unstable correspondence or limited available gene numbers. Overall, the results demonstrate that the tetraploid *G. pentaphyllum* exhibits clear subgenomic differentiation across most chromosomes, suggesting an allopolyploidy origin of the tetraploid.

To investigate genomic architecture between tetraploid SN and diploid NN, we performed systematic whole-genome synteny analysis of pairwise haplotype genomes (Figure S4B). All chromosomal pairs demonstrated extensive collinearity, indicative of evolutionarily conserved genome organisation. Nevertheless, localised structural rearrangements and syntenic disruptions were identified, revealing regions of structural variation. Whole-genome alignment further corroborated these findings, with most chromosomal segments exhibiting continuous diagonal patterns characteristic of conserved gene order, while deviations and discontinuities highlighted structural divergences. These aberrant regions correspond to rearrangements, insertions, or deletions previously annotated by JCVI. Collectively, this analysis reveals both global structural conservation and localised genomic differentiation between tetraploid and diploid *G. pentaphyllum*, providing insights into their evolutionary divergence and potential phenotypic variation.

2.4 | Identifying the Biosynthetic Genes Involved in Dammarene-Type Gypenoside Biosynthesis

To identify genes implicated in the biosynthetic pathway of tetracyclic triterpenoids in *G. pentaphyllum* (Figure 3A), we conducted transcriptome sequencing across multiple tissues from two geographically distinct ecotypes (NN and SN) and assessed differential expression profiles. Candidate genes encoding catalytic enzymes for upstream tetracyclic triterpenoid biosynthesis were identified through homology-based protein sequence alignment. This analysis revealed genes encoding four key terpenoid biosynthetic enzymes: five FPPS genes, two SS genes, two SE genes, and one DS gene. FPPS1 and FPPS2 were predominantly expressed in SN-R, NN-L, and NN-R, while FPPS3-5 showed high expression levels in SN-R, SN-S, and SN-L. SS1 and SE1 exhibited root-specific expression in SN, while SS2 and SE2 displayed broad expression in tissues of HN. DS was highly expressed in SN-R and NN-R/S/L. The DS gene was highly expressed in HN roots and all examined in tissues of NN. Notably, the identified DS gene exhibited near-complete sequence homology (> 98%) with the functionally characterised GpOSC1 and conserved root-enriched expression—particularly in adventitious roots (AR)—recapitulating previously reported GpOSC1 expression patterns (Figure 3B) (Yun et al. 2024).

To further investigate hydroxylases in the ginsenosides and gypenoside LVI biosynthesis of *G. pentaphyllum*, we systematically identified CYP450 genes through HMMER profiling (PF00067) and local BLAST analysis. Following removal of incomplete sequences, 96 high-confidence CYP450 candidates exhibiting expression (FPKM > 20) in SN ecotypes were retained (Figure S4C). Phylogenetic reconstruction incorporating functionally annotated CYP450s from related species revealed no clustering with known *Panax* PPDS homologues or members of the CYP716 subfamily (Figure S5). This suggests hydroxylases functionality in *G. pentaphyllum* may derive from a novel CYP450 family distinct from CYP716, warranting further characterisation. We prioritised highly expressed candidates (FPKM > 20) across SN tissues for validation. Among the 96 CYP450s, 25 exhibited significantly elevated expression in SN relative to NN ecotypes. Correlation analysis between these 25 CYP450s and seven quantified triterpenoid compounds identified GpCYP88-17, GpCYP88-18, and GpCYP88-25 as showing strong positive correlations ($r > 0.7$) with ginsenosides Rb1, Rd, Rb3, F2, Rg5, CK and Rg3. Correlations with Rb1, Rd, Rb3, F2 and Rg5 were statistically significant ($p < 0.01$) (Figure 3C). Based on these associations, these CYP88 family members were selected as prime candidates for functional validation.

2.5 | Functional Verification of the CYP450 for C12-Hydroxylation and C2-Hydroxylation of Dammarenediol II

Given the demonstrated superiority of plant heterologous expression systems for terpenoid biosynthesis, we employed *Nicotiana benthamiana* for transient co-expression assays. Three *G. pentaphyllum* CYP450 candidates (including *GpCYP88-18*) were co-infiltrated with *Fusarium* squalene epoxidase gene (*FusSQE*) and *GpDS* via *Agrobacterium*-mediated transformation (Figure 4A). Three days post-infiltration, squalene substrate was introduced to assess enzymatic activity. LC–MS analysis of the *N. benthamiana* leaf extracts revealed that only the samples co-expressing *FusSQE*, *GpDS*, and *GpCYP88-18* showed a distinct peak at 5.287 min, exhibiting mass fragmentation patterns and retention time identical to both the positive control (the product of co-expressing *FusSQE*, *GpDS*, and *P. quinquefolius PPDS*) and authentic PPD standard (Figure 4B,C). In contrast, co-expressing *FusSQE*, *GpDS*, GpCYP88-17 or GpCYP88-25 does not yield the target product (Figure S6). These results demonstrate GpCYP88-18 possesses specific PPDS activity, converting squalene-derived intermediates to PPD. Notably, we also detected 2α-OH-PPD (RT = 4.428 min) exclusively in the samples (co-expressing *FusSQE*, *GpDS* and *GpCYP88-18*). The MS fragments of the peak at 4.428 min showed high consistency with the MS profile of the 2α-OH-PPD standard. The absence of a corresponding peak in negative controls at this retention time (4.428 min) confirms the enzymatic origin of the product. Based on these findings, the results indicated that GpCYP88-18 exhibits bifunctional catalytic activity, catalysing the conversion of dammarenediol II to PPD and its subsequent hydroxylation at C-2 to yield 2α-OH-PPD (Figure 4D,E).

Furthermore, we functionally validated GpCYP88-18 activity in *Saccharomyces cerevisiae*. LC–MS analysis detected a novel chromatographic peak at 5.142 min exclusively in the BY-DS strain co-transformed with pCEV-Leu-AtCPR1 and pCEV-Trp-GpCYP88-18, exhibiting retention time and fragmentation patterns congruent with authentic PPD standard (Figure 5A). The absence of this peak in negative controls confirmed GpCYP88-18 catalyses the conversion of dammarenediol-II to PPD (Figure 5B,C). Consistent with tobacco transient assays, we also identified 2α-OH-PPD production in this heterologous system (Figure 5D,E). These results demonstrate GpCYP88-18 possesses bifunctional catalytic activity, first catalysing the C-12 hydroxylation of dammarenediol II to yield PPD and then the C-2 hydroxylation to produce 2α-OH-PPD. Formally designated CYP88AB3 by the International CYP450 Nomenclature Committee, the enzyme has been confirmed as a member of the CYP88 family. Significantly, GpCYP88AB3 is the first enzyme characterised with both C-12 hydroxylation and C-2 hydroxylation activity in *G. pentaphyllum*. As such, it bridges a pivotal catalytic gap in the biosynthesis of ginsenosides and gypenoside LVI, enabling their heterologous production.

To investigate the function of *GpCYP88AB3* in planta, we used Virus-Induced Gene Silencing (VIGS) to suppress its expression in *G. pentaphyllum*. Using *PDS* as a positive control, we observed that the silencing of *GpPDS* resulted in an albino-yellowing phenotype in the leaves of *G. pentaphyllum* seedlings (Figure 6A). The qRT-PCR analysis confirmed that *GpPDS* expression was significantly suppressed, indicating the successful establishment of the VIGS system in *G. pentaphyllum*. Subsequently, suppression of *GpCYP88AB3* expression showed that the plants grew normally compared with the negative control, and expression analysis confirmed significant downregulation of the target gene in the experimental group (Figure 6B,C). Further metabolic profiling of experimental and control plants revealed a significant decrease in the content of main triterpenoid

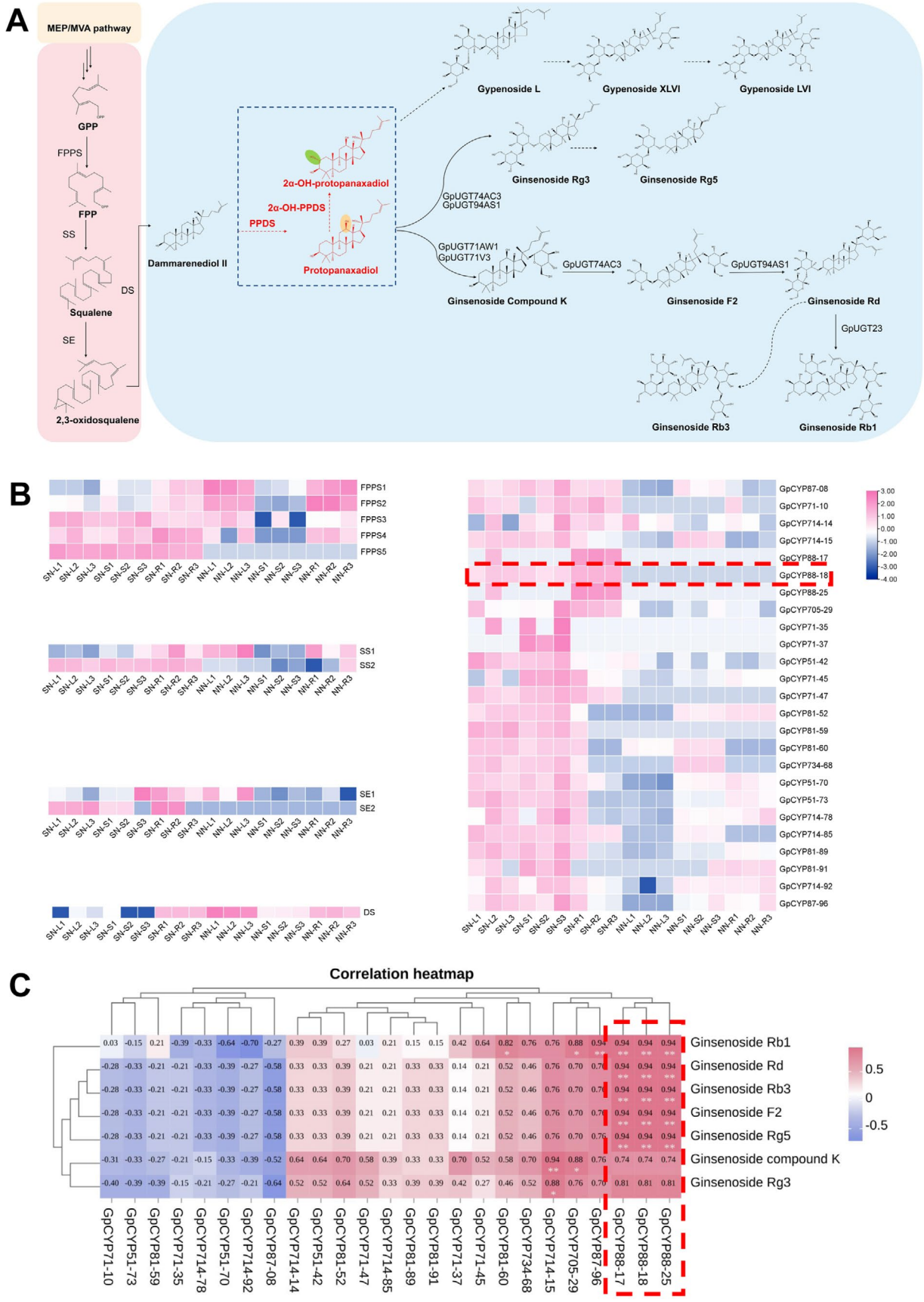

FIGURE 3 | Hypothetical ginsenosides and gypenoside LVI biosynthetic pathway, and identification of candidate P450 genes in *G. pentaphyllum*. (A) The biosynthetic pathways for ginsenosides and gypenoside LVI in *G. pentaphyllum*. (B) The expression levels of key genes in the ginsenosides and gypenoside LVI biosynthetic pathway. (C) The high expression levels of candidate CYP450 genes in SN. ** represents $p < 0.01$, and * represents $p < 0.05$.

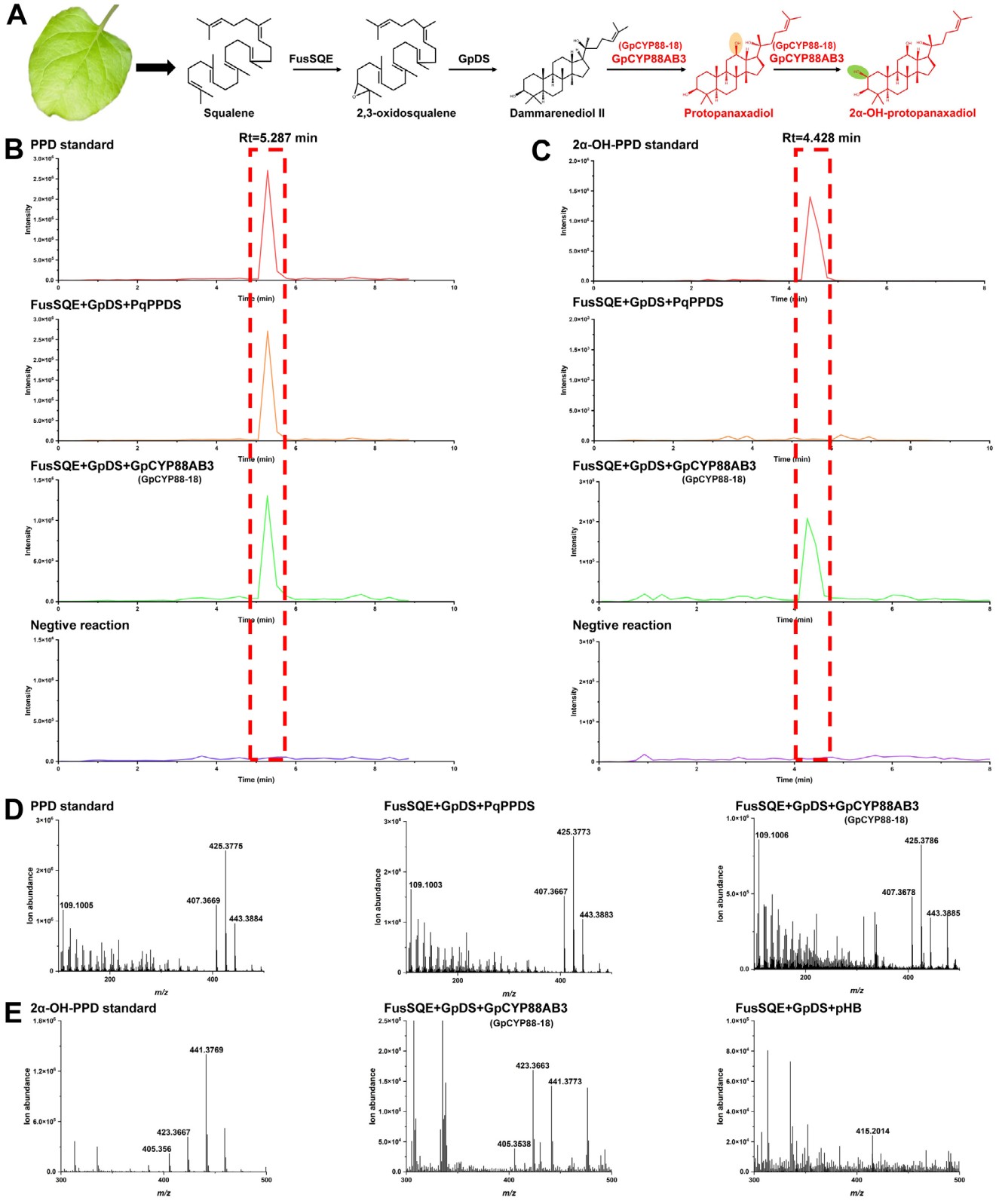

FIGURE 4 | Functional verifications of GpCYP88AB3 in *N. benthamiana*. (A) FusSQE, GpDS, and GpCYP88AB3 were co-expressed in tobacco. (B) The Extracted PPD TIC chromatogram of PPD standard, experimental group (FusSQE + GpDS + GpCYP88AB3), positive control group (FusSQE + GpDS + PqPPDS) and negative control group. (C) The Extracted 2α-OH-PPD TIC chromatogram of 2α-OH-PPD standard, experimental group (FusSQE + GpDS + GpCYP88AB3) and negative control group (FusSQE + GpDS). (D) MS profiles of PPD standard, experimental group and positive control. (E) MS profiles of 2α-OH-PPD standard, experimental group and negative control.

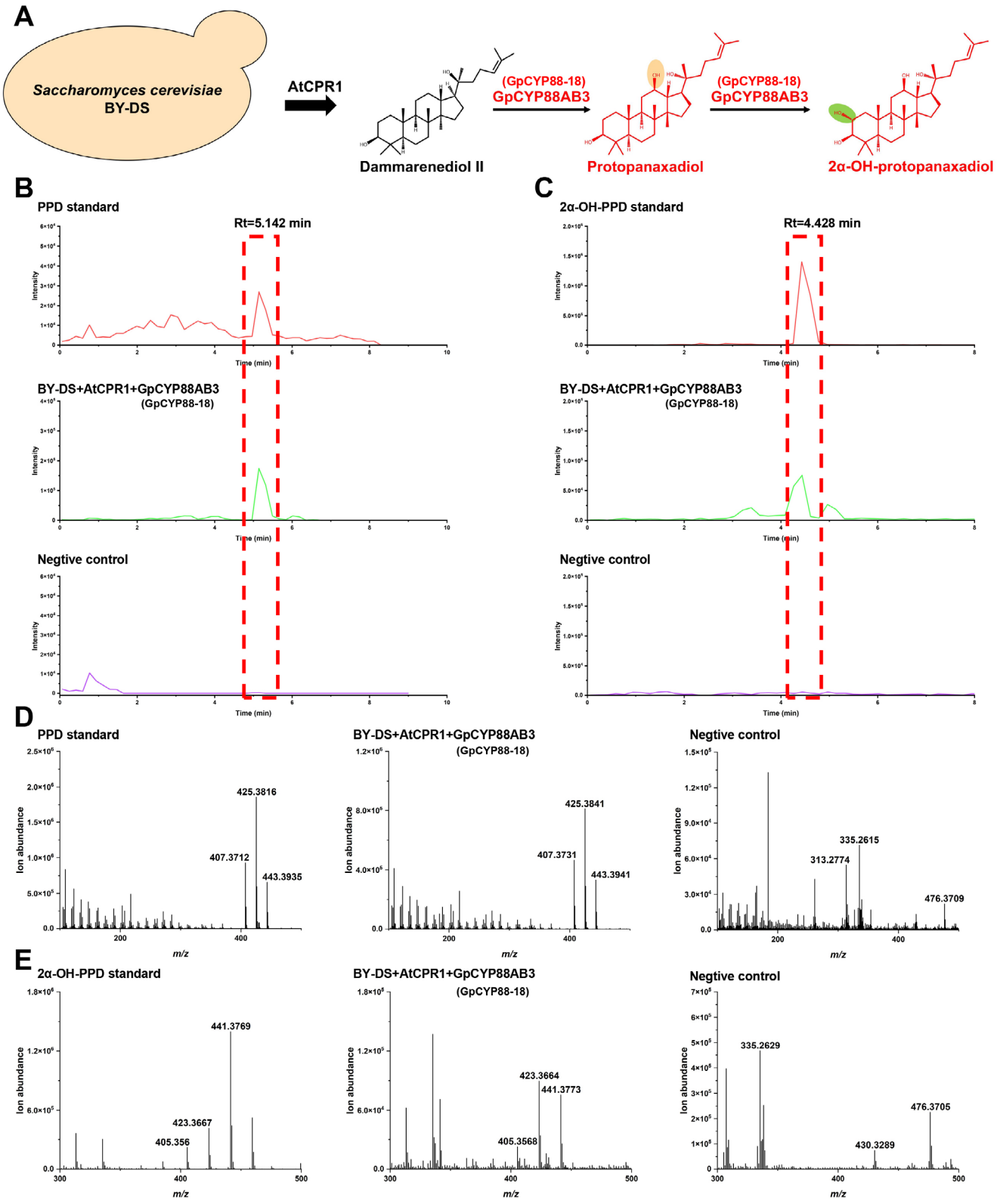

FIGURE 5 | Functional verifications of GpCYP88AB3 in *S. cerevisiae*. (A) AtCPR1 and GpCYP88AB3 were co-expressed in *S. cerevisiae*; (B) the extracted PPD TIC chromatogram of PPD standard, experimental group (BY-DS + AtCPR1 + GpCYP88AB3), and negative control group (BY-DS + AtCPR1); (C) the extracted 2α-OH-PPD TIC chromatogram of 2α-OH-PPD standard, experimental group (BY-DS + AtCPR1 + GpCYP88AB3), and negative control group; (D) MS profiles of PPD standard, experimental group and negative control. (E) MS profiles of 2α-OH-PPD standard, experimental group and negative control.

saponins including PPD, CK, Rb1, XLVI and LVI. The upstream precursor of GpCYP88AB3, Dammarenediol-II, showed significant accumulation in the experimental group. The above results suggested that silencing *GpCYP88AB3* disrupts the downstream biosynthesis of PPD type and 2α-OH-PPD type triterpenoid saponins (Figure 6D–I).

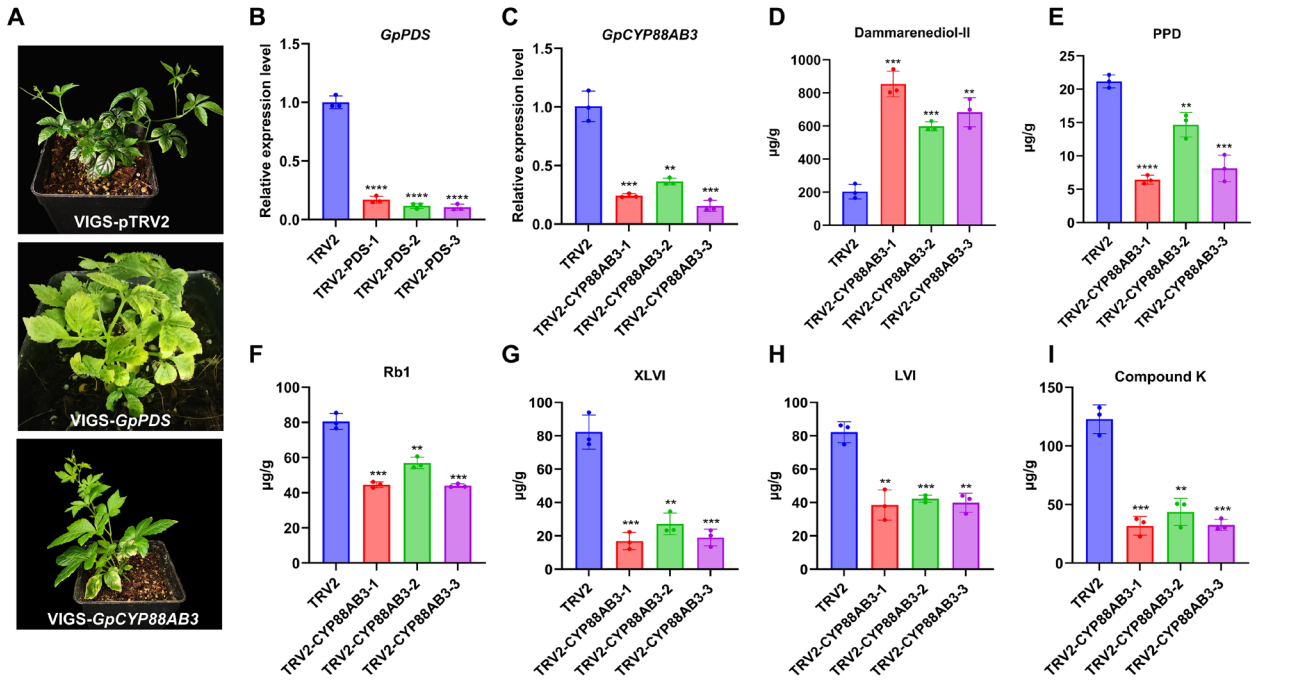

FIGURE 6 | The VIGS assay of *GpCYP88AB3* in *G. pentaphyllum*. (A) The plant phenotypes of positive, experiment and negative groups. (B, C) The *GpPDS* and *GpCYP88AB3* gene expression levels of different groups. (D–I) The content of Dammarenediol-II, PPD, CK, Rb1, XLVI and LVI in experiment and negative groups. ** indicates $p < 0.01$, *** indicates $p < 0.001$, and **** indicates $p < 0.0001$.

2.6 | The Evolution of the CYP88 Family in SN *G. pentaphyllum* and the Emergence of GpCYP88AB3 Activities

Furthermore, to elucidate the evolutionary dynamics of CYP88 enzymes within Cucurbitaceae, we conducted genome-wide identification across 11 representative species, identifying 3839 CYP genes. Phylogenetic reconstruction based on full-length amino acid sequences resolved these CYPs into five major clans (Figure 7A). Notably, the experimentally validated CYP88 family formed a monophyletic clade within the CYP85 clan (Figure 7A, red branch), confirming its phylogenetic positioning. This contrasts with *Panax ginseng* PPDS (PgPPDS), which resides in the CYP716 family, indicating that the two PPDSs emerged independently by convergent evolution.

To elucidate the evolutionary origin of *GpCYP88AB3* within the Cucurbitaceae CYP88 family, we conducted a targeted phylogenetic analysis (Figure 7B). This resolved two distinct sister clades: one comprising diverse Cucurbitaceae CYP88s (collapsed as 'Other Cucurbitaceae CYP88s'), and a monophyletic *Gynostemma*-specific clade containing exclusively *Gynostemma* CYP88 homologues. Crucially, our functionally characterised GpCYP88-18 (red highlight, Figure 7B) clustered within this *Gynostemma*-specific clade, which comprises homologues from both diploid and tetraploid *G. pentaphyllum*. Notably, tetraploid *G. pentaphyllum* exhibited a pronounced expansion of CYP88 genes relative to its diploid counterpart, implying lineage-specific duplication events following polyploidisation that drove functional diversification.

Microsynteny analysis across the 11 Cucurbitaceae species revealed that CYP88 members within the 'Other Cucurbitaceae CYP88s' clade share conserved collinearity across all 11 species (including SN and NN), confirming the presence of an ancient, conserved CYP88 lineage in Cucurbitaceae (Figure S7). In contrast, the Gynostemma-specific CYP88 clade lacks such microsyntenic conservation, supporting its origin as a *Gynostemma*-specific evolutionary innovation. The sister-clade relationship between the conserved 'Other Cucurbitaceae CYP88s' and the *Gynostemma*-specific clade indicates a specific duplication event of CYP88 in a *Gynostemma* ancestor. Notably, within the *Gynostemma*-specific clade, SN exhibits a significant expansion of CYP88 members, which is likely driven by both tandem duplication and ectopic duplication events. This proliferation of CYP88 genes in SN may have facilitated neofunctionalisation, potentially endowing these enzymes with novel catalytic activities that underpin the species-specific biosynthesis of ginsenosides—insenosides endowing the Cucurbitaceae species and NN. As the gene encoding GpCYP88AB3 shows no synteny to other CPY88s and is located at a locus with our nearly CPY88s as well (Figure 7C), it is likely that GpCYP88AB3 arose by ectopic duplication in the SN (Huang et al. 2024; Xue et al. 2023).

3 | Discussion

As a significant member of the Cucurbitaceae family, *G. pentaphyllum* possesses substantial medicinal value and health-promoting properties (Li et al. 2025). Chromosome-level genome assembly strategies have currently been applied to numerous medicinal plants, such as *Lonicera japonica* (Yin et al. 2023), *Dendrobium officinale* (Niu et al. 2021), *Isatis indigotica* (Kang et al. 2020), and *Callicarpa nudiflora* (Huang

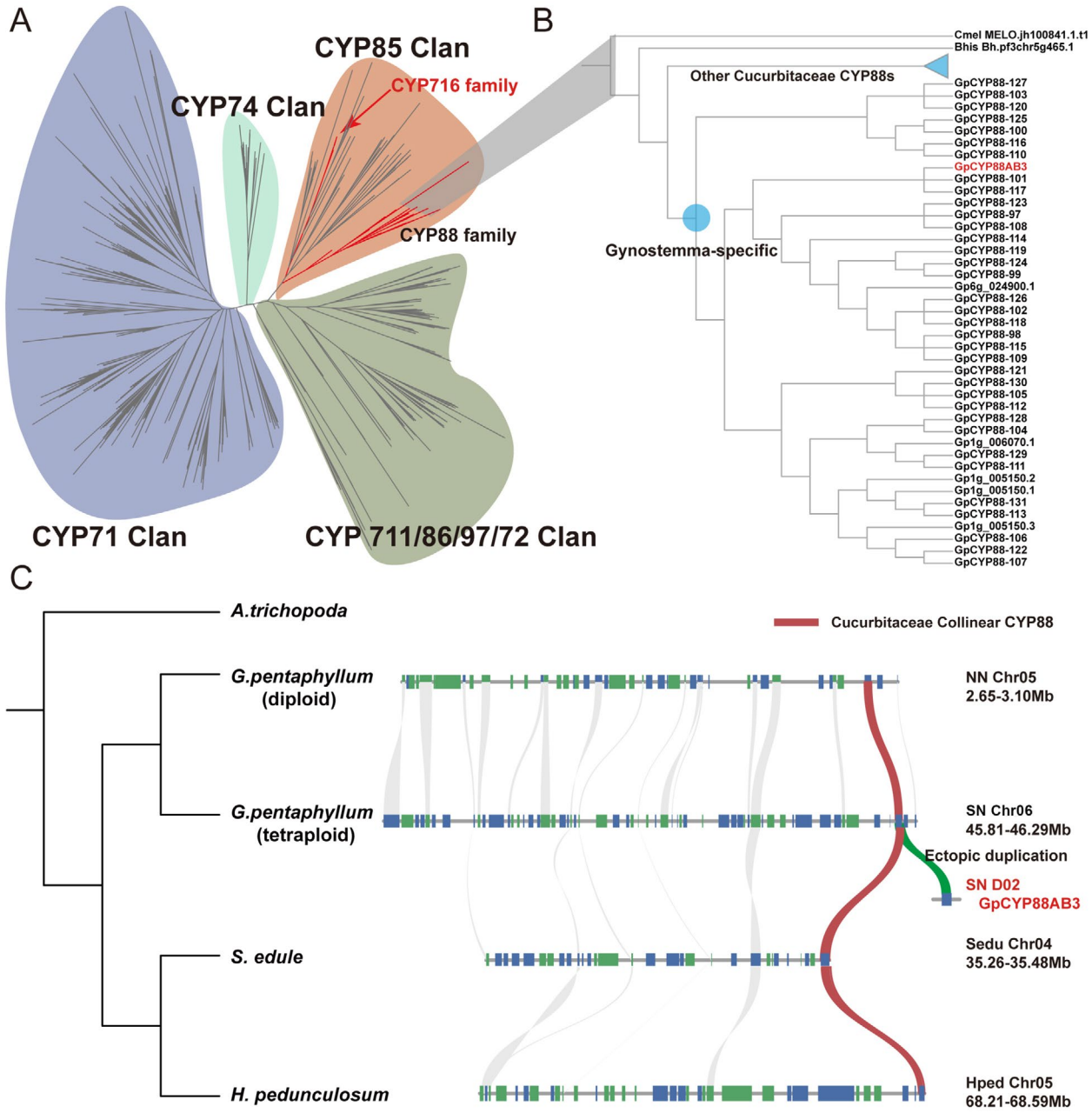

FIGURE 7 | Phylogenetic analysis of Cucurbitaceae CYP genes and lineage-specific expansion of the CYP88 subfamily. (A) Maximum-likelihood phylogenetic tree of Cucurbitaceae CYPs. Clans are colour-coded: CYP71 Clan (blue), CYP74 Clan (light green), CYP85 Clan (orange), CYP88 Clan (red), and CYP711/86/97/72 Clan (green). The CYP88 subfamily is nested within the CYP85 Clan (red branch). (B) Expanded view of the CYP88 subfamily phylogenetic tree. The grey collapsed triangle represents the 'Other Cucurbitaceae CYP88s' clade (conserved across 11 Cucurbitaceae species). The sister 'Gynostemma-specific CYP88 clade' (no microsynteny) includes both NN and SN *G. pentaphyllum* sequences, with the functionally validated gene *GpCYP88AB3* highlighted in red. (C) Conserved microsynteny of CYP88 genes in Cucurbitaceae, with *GpCYP88AB3* located on chromosome D02 of SN versus conserved CYP88s on chromosome 06. The green lines indicate ectopic duplication, while the red lines represent collinear homologous genes.

et al. 2025), to elucidate the biosynthetic pathways of bioactive compounds and explore genomic features underlying species evolution. Notably, *G. pentaphyllum* represents the sole non-Araliaceae species capable of ginsenoside production (Yuan et al. 2024). However, previous studies have shown that only some natural populations of *G. pentaphyllum* contain ginsenosides, and the specific formation mechanism is still unclear. Natural populations of *G. pentaphyllum* from Suining (SN) and Nanning (NN) in China show significant divergence in the content of valuable dammarane-type ginsenosides and gypenoside LVI (Chen, Yao, et al. 2023; Liang et al. 2025; Wang et al. 2020). Therefore, we selected a ginsenoside-rich natural population from Suining, Hunan Province to study the formation mechanism. Our study successfully constructed a chromosome-level reference genome from Suining, Hunan Province and revealed that it is a tetraploid *G. pentaphyllum*. This high-quality chromosome-scale assembly established a crucial foundation for investigating chromosome evolution

and WGD events within this species and the broader genus (Chen et al. 2023). This resource provides an unprecedented resolution for studying subgenome evolution in *G. pentaphyllum* and the diversity of ginsenosides and gypenoside LVI biosynthetic pathways.

Gene family expansions and contractions constitute key drivers of genetic innovation (David et al. 2025). *GpCYP88AB3* originated through such expansion, warranting future investigations into the molecular evolutionary mechanisms and functional diversification of these modified gene families. Whole-genome duplications (WGDs) critically facilitate genus-wide species diversification, morphological innovation, and functional novelty. Prior studies indicate four ancestral WGD events in Cucurbitaceae, including an early cucurbit-common tetraploidisation (CCT) event following the core-eudicot γ-WGT (Guo et al. 2020; Wang et al. 2018). Notably, our analyses detected no CCT signature in either *G. pentaphyllum* genome (diploid or tetraploid), but identified a population-specific WGD in the tetraploid SN lineage post-dating γ-WGT. These hierarchically distinct WGD events establish layered evolutionary frameworks, with the recent tetraploid-specific WGD potentially enabling metabolic pathway expansion for specialised compound biosynthesis.

Chromosome evolution analysis revealed that tetraploid *G. pentaphyllum* exhibits stable subgenome pairing patterns at the chromosomal level, and both the species tree topologies and quartet statistics showed clear signals of allopolyploidy. This conclusion is consistent with many well-studied allopolyploid species. For instance, in *Brassica napus*, the A and C subgenomes of different origins can also be clearly distinguished in chromosomal collinearity and phylogenetic analyses, confirming allopolyploidisation in the Brassica ancestor (Chen et al. 2021). Similarly, in strawberry (*Fragaria* × *ananassa*) (Edger et al. 2019) and wheat (*Triticum aestivum*) (Feldman and Levy 2012), the formation of allopolyploids is often accompanied by unequal conservation and functional differentiation between subgenomes. Our results indicate that the evolutionary trajectory of the tetraploid *G. pentaphyllum* is similar to these patterns, suggesting that an allopolyploidy origin for the tetraploid *G. pentaphyllum*. It is noteworthy that a few chromosomes (such as chr08, chr09 and chr11) exhibited ambiguous or mixed phylogenetic signals, which may be related to chromosomal rearrangements, gene loss, or introgression between subgenomes. Similar phenomena have also been reported in strawberry and wheat, usually reflecting secondary evolutionary processes following polyploidisation. This suggests that tetraploid *G. pentaphyllum* not only originated through an allopolyploidisation event but may also have experienced subsequent variations between subgenomes.

Hydroxylation constitutes the critical catalytic step for ginsenoside and gypenoside LVI biosynthesis in *G. pentaphyllum* (Liang et al. 2019). PPD-type saponins are distributed in both *Panax* and *Gynostemma*. However, the biosynthesis of PPT-type saponins is confined to *Panax*, and that of 2α-OH-PPD derivatives is specific to *Gynostemma*. Notably, PPT and 2α-OH-PPD are structural isomers sharing the same molecular formula. The biosynthetic pathway of PPD-type saponins initiates with PPDS, which produces the dammarane backbone PPD—subsequently glycosylated by UGTs to form PPD-type ginsenosides (e.g., CK, F2 and Rg3). This biosynthetic divergence is determined at the hydroxylation step. In *Panax*, PPTS (CYP716 subfamily) catalyses C-6 hydroxylation of PPD to yield PPT, the precursor of ginsenosides such as Rh1, F1 and Rg1. Conversely, in *G. pentaphyllum*, the specific enzyme GpCYP88AB3 mediates C-2α hydroxylation, generating 2α-OH-PPD—the direct precursor of gypenoside LVI.

Within Araliaceae, PPDS functionality has been functionally validated in multiple *Panax* species: PgCYP716A47 (*P. ginseng*), PnCYP716A47 (*P. notoginseng*) and PqCYP716A47 (*P. quinquefolius*) (Cun et al. 2024; Han et al. 2011; Wang et al. 2014). Similarly, PPTS activity is mediated by homologous CYP716A subfamily members: PgCYP716A53v2 (*P. ginseng*), PnCYP716A53v2 (*P. notoginseng*) and PqCYP716A53v2 (*P. quinquefolius*). These evolutionarily conserved PPDS and PPTS enzymes share high sequence homology within the CYP716A subfamily. Crucially, although this role is evolutionarily conserved in *Panax*, we did not identify any CYP716A homologues among the 96 high-confidence CYP450 candidates from *G. pentaphyllum*. This absence prompted further investigation of the CYP450 family in *G. pentaphyllum*. We define GpCYP88AB3 as the enzyme that catalyzes the sequential regiospecific hydroxylation of dammarenediol-II at C-12 and C-2α to yield 2α-OH-PPD. Strikingly, GpCYP88AB3 exhibits less than 30% sequence identity to known *Panax* PPDS orthologs (Figure S8), thus providing a clear case of convergent evolution of this function in distantly related CYP450 subfamilies. These results indicate an independent origin for PPDS enzymes in *G. pentaphyllum* and *Araliaceae*. Crucially, GpCYP88AB3 exhibits catalytic promiscuity enabling consecutive hydroxylations, whereas *Panax* enzymes (PgPPDS and PgPPTS) perform single oxidations. These findings reveal *G. pentaphyllum* evolved a bifunctional P450 system through subfunctionalisation—representing a divergent molecular evolutionary trajectory from the Araliaceae lineage.

In summary, this study presents the first high-quality tetraploid reference genome for *G. pentaphyllum* (SN type) and provides strong evidence supporting its allopolyploid origin. Our analyses reveal variations in chromosomal conservation and rearrangement, offering unprecedented insights into the genomic architecture of *G. pentaphyllum* and the biosynthetic machinery responsible for gypenoside production. By integrating comparative genomics, metabolomics, and transcriptomics, we systematically identified candidate CYP450 genes in *G. pentaphyllum* and functionally characterised the enzymatic activity of *GpCYP88AB3* using dual validation systems: in vivo assays in *N. benthamiana* and in vitro reconstruction in *S. cerevisiae*. Furthermore, the role of *GpCYP88AB3* was confirmed in *G. pentaphyllum* through VIGS. These results shed light on genome evolution and species formation within the Cucurbitaceae family, and establish a foundation for future investigations into subgenome functional differentiation, gene expression dominance, and adaptive evolution. This work also provides crucial groundwork for elucidating the complete gypenoside biosynthetic pathway, developing platforms for heterologous production of these high-value compounds, and supplying genetic resources for molecular-assisted breeding and genetic improvement of *G. pentaphyllum*.

4 | Materials and Methods

4.1 | Plant Material

The SN *Gynostemma pentaphyllum* used for genome sequencing was collected from Yuanjialing, Shaoyang, Hunan Province (cultivated variety). The different tissues of SN, including leaves, stems and roots, were collected for sequencing and metabolomic analyses. These tissues were thoroughly washed three times with ultrapure water to remove external contaminants. Subsequently, the leaves were frozen in liquid nitrogen and stored at −80°C until DNA and RNA extraction. For metabolomic analyses of NN *G. pentaphyllum*, the different tissues of leaves, stems, and roots were collected from Nanning, Guangxi Province, China. For the genome survey of SN, tender leaves were collected and utilised. High-quality genomic DNA was extracted from the fresh leaf tissues of *G. pentaphyllum* and used to construct libraries for Illumina, PacBio, and Hi-C sequencing.

4.2 | Chemicals

Reference standards for Ginsenoside Rb1, Ginsenoside Rd, Ginsenoside Rb3, Ginsenoside compound K, Ginsenoside F2, Ginsenoside Rg3, Ginsenoside Rg5, Protopanaxadiol, Gypenoside XLVI and Gypenoside L were purchased from Chengdu DeSiTe Biological Tech-nology Co. Ltd. (China). The 2α-OH-protopanaxadiol and Gypenoside LVI standards were provided by Professor Yin Zhiqi from China Pharmaceutical University.

4.3 | Metabolome Profiling of *G. pentaphyllum* From SN and NN

Metabolites were extracted from different tissues of SN and NN using 80% methanol, which tissues were extracted by ultrasonic extraction equipment (KH-300SP ultrasonic extraction equipment, China) for 3 times, 30 min each time. The supernatant was collected after centrifugation at 13 000 rpm for 10 min and then went through a 0.22-μm filtration membrane before analysis. Samples were analysed by high-performance liquid chromatography–ultraviolet–mass spectrometry (HPLC–UV–MS/MS) using an AB Sciex TripIeTOF X500 LC/MS/MS instrument equipped with an ExionLC UHPLC unit (AB SCIEX, USA). Samples of 5 μL were injected onto an InfinityLab Poroshell 120 EC-C18 column (Agilent, 100 × 2.1 mm, 1.9 μm). The column temperature was held at 42°C, and the flow rate was held constant at 0.35 mL min^{-1}. A gradient using 0.1% *v*/*v* formic acid–acetonitrile as Buffer A and 0.1% *v*/*v* formic acid–water as Buffer B was run as follows: 12%–27% Buffer B from 0 to 8 min, 27%–40% Buffer B from 8 to 23 min, 40%–74% Buffer B from 23 to 26 min, 74%–100% Buffer B from 26 to 29 min, 100%–12% Buffer B from 29 to 32 min, 12% Buffer B from 32 to 35 min. For mass spectrometry, the ionisation mode was by electrospray ionisation (ESI), the acquisition mode was information-dependent acquisition (IDA), and negative ion mode was used. The MS parameters were curtain gas, 35 psi; ion spray voltage, −4500 V; gas temperature, 550°C; ion source gas 1 and gas 2, 55 psi; collision energy, 40 ± 20 V; mass range, 50–1700 m/z; and acquisition rate, 20 frames s^{-1}. Chromatograms and mass spectra were analysed using AnalystTF Software 1.7 (Sciex). Data processing was performed with PeakView 2.2 software (Sciex) and MasterView 1.3 software (Sciex). Each sample was analysed in triplicate.

The following standards including Ginsenoside Rb1, Ginsenoside Rd, Ginsenoside Rb3, Ginsenoside compound K, Ginsenoside F2, Ginsenoside Rg3, Ginsenoside Rg5, and Protopanaxadiol with different concentrations were used to establish linear equations (Table S8), and the content of components in the samples was determined according to the external standard method.

4.4 | Genome Assembly and Scaffolding

The initial assembly of the *G. pentaphyllum* genome was performed with hifiasm v0.19.5-r587 using PacBio HiFi reads. The raw contigs were polished with NextPolish v1.4.1, also leveraging the HiFi dataset, to correct residual base-level errors. To remove redundant haplotigs and collapse heterozygous contigs, we applied Purge_haplotigs v1.1.2 to the polished assembly. For chromosome-scale scaffolding, Hi-C data were processed with ALLHIC v0.9.12 to group and scaffold the error-corrected contigs. The resulting interaction matrices were converted into the .hic format via Juicebox, and manual curation within the Juicebox Assembly Tools interface ensured correct contig ordering and orientation to produce the final chromosome-level assembly.

4.5 | Repeat Annotation and Gene Prediction

Repetitive elements in the *G. pentaphyllum* genome were discovered de novo using RepeatModeler v2.0.2 with default parameters, generating a species-specific repeat library (consensi.fa.classified). This library served as the input for RepeatMasker v4.1.0 to annotate and mask repetitive regions across the genome. Gene structures were annotated by integrating ab initio prediction, homology-based evidence, and transcriptome data. Specifically, masked genomic sequences were processed with BRAKER3 v3.0.3, which combined AUGUSTUS and GeneMark-ES ab initio predictions, protein homology from closely related species, and RNA-seq alignments to produce a non-redundant, high-confidence gene set.

4.6 | Assembly and Annotation Quality Assessment

To evaluate the completeness of both the assembly and the gene annotation, we ran BUSCO v4.1.2 in genome and protein modes against the embryophyta_odb10 database. A completeness score (Complete BUSCOs) exceeding 90% was considered indicative of a highly intact and well-annotated genome.

4.7 | Comparative Phylogenomic Analysis

We retrieved genome assemblies and GFF3 annotation files for 19 phylogenetically representative angiosperms—spanning monocots (*Oryza sativa*), eudicots (*Saururus chinensis, Vitis vinifera,*

Solanum tuberosum, Solanum lycopersicum, Peucedanum praeruptorum, Aralia elata, Panax ginseng, Panax japonicus, Populus trichocarpa, Arabidopsis thaliana, Gynostemma pentaphyllum (diploid), *Sechium edule, Herpetospermum pedunculosum, Cucurbita maxima, Cucurbita moschata, Citrullus lanatus, Benincasa hispida, Cucumis melo, Cucumis hystrix and Cucumis sativus*) and the basal angiosperm *Amborella trichopoda*. For each species, we extracted the longest transcript per gene and retained its corresponding protein sequence. Orthologous single-copy gene families were identified using OrthoFinder v2.5.2. Multiple sequence alignments of these conserved genes were generated with ClustalW v2.0.10, and concatenated into a supermatrix. We reconstructed the species phylogeny under the maximum-likelihood framework in IQ-TREE v2.2.0-beta, employing ultrafast bootstrap approximation for branch support. Gene family expansion and contraction dynamics were inferred with CAFE5, while divergence times were estimated using MCMCTree calibrated with fossil constraints obtained from TimeTree.

4.8 | Identification of Polyploidisation Events

To systematically characterise polyploidy in *G. pentaphyllum*, we combined synonymous substitution rate (Ks) distribution analysis with whole-genome synteny assessment. First, we processed the raw protein sequences using a custom Python script to retain only the longest transcript per gene. Next, we identified paralogous gene families in *G. pentaphyllum* and related species via the MCL algorithm implemented in WGD v1.1. Intra- and inter-species Ks values were then computed using the 'ksd' module, while collinear blocks and their median Ks distributions were delineated with the 'syn' module. We visualised the resulting Ks spectra in R to distinguish whole-genome duplication (WGD) signatures from speciation-derived divergence. To further validate genome-wide collinearity and block identification, we employed WGDI v0.6.2, mapping Ks values onto syntenic anchor pairs to resolve both ancient and recent WGD events. By integrating the Ks distribution profiles, the spatial arrangement of homologous blocks, and phylogenetic context, we inferred the timing and lineage of WGD occurrences in the *G. pentaphyllum* genome.

4.9 | Chromosome Evolution Analysis

We employed a combined collinearity and phylogenetic approach to analyse the origin of tetraploid *G. pentaphyllum*. First, MCScanX was used to conduct whole-genome collinearity analysis between the diploid reference genome (R) and the four subgenomes (A, B, C and D) of the tetraploid. Homologous gene relationships were established using diamond blastp, and collinear blocks were identified at the chromosomal level. Subsequently, homologous gene sets containing all copies from both the diploid and tetraploid were extracted within each chromosome, and single-gene phylogenetic trees were constructed. Finally, all single-gene trees were integrated using ASTRAL (V5.6.3) to generate chromosome-level species trees, thereby determining the phylogenetic relationships among subgenomes (Zhang et al. 2018).

4.10 | Transcriptome Profiling of Different Tissues of SN and NN

Total RNA was extracted from different tissues (leaves, stems and roots with three biological repeats) using Trizol reagent kit (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's protocol. The mRNA from all samples was purified using the TruSeq RNA Sample Prep Kit (Illumina), followed by sequencing on an Illumina HiSeq 4000 sequencer. Raw RNA reads were filtered and trimmed to yield clean reads by Trimmomatic (v0.39) and FastQC (v0.11.9), and then these highquality reads were mapped to the draft reference genome by HISAT2. Gene expression was calculated with Fragments Per Kilobase of exon model per Million mapped fragments (FPKM). The differential patterns of gene expression were analysed with count number by DESeq2 using a model based on the negative binomial distribution. DESeq2 was used to normalise gene expression (BaseMean) in each sample, and differentially expressed genes (DEGs) were identified for each pair of compared groups by using p adj. (adjusted p) < 0.05 as the threshold. The gene heatmap figures were produced by using TBtools with Heatmap Illustrator function.

4.11 | Screening of the P450 Genes Involved in Ginsenoside and Gypenoside LVI Biosynthesis

To screen the P450 genes involved in the formation of ginsenoside and gypenoside LVI, we screened the functional annotation information of the genome and selected the protein sequences of the identified genes in *G. pentaphyllum*. These sequences were then subjected to BLASTPv2.11.0 (E-value $< 1e^{-5}$) analysis. We used HMMER (version 3.3.2) and Local Blast to scan all the predicted CYP450 in *G. pentaphyllum*, retraining those containing the "Cytochrome P450 family" (PF01040) domain. Candidate P450 genes were screened through phylogenetic analysis along with CYP450 genes from *Arabidopsis thaliana* and protein sequences with verified functions in other plants.

4.12 | Evolution of CYP450 Genes

To analyze the evolution of CYP proteins, we performed a combined search using BLAST and HMMER (PF00067) against proteomes of 11 Cucurbitaceae species, merged the results, and filtered out sequences shorter than 200 amino acids, with conserved CYP domains further verified by NCBI CDS-Search. The resulting CYP protein sequences were aligned using MAFFT v7.49 followed by manual refinement, and a maximum likelihood phylogenetic tree was constructed using IQ-TREE v2.0.7 under default parameters. For synteny analysis, we identified CYP gene collinearity blocks across the 11 species using JCVI and visualised the syntenic relationships to generate final plots.

4.13 | Functional Analysis of GpCYP88AB3 in *N. Benthamiana*

Based on the transcriptional expression changes of the CYP450 gene family, candidate genes were screened using an FPKM > 20 threshold for functional validation in *N. benthamiana*.

The open reading frame (ORF) of the candidate CYP450 was inserted into the pHB binary expression vector via double digestion with the restriction enzymes *Hin*dIII and *Pst*I. The construct was then transformed into *Agrobacterium tumefaciens* GV3101 using the freeze–thaw method, and positive transformants were selected and verified by agarose gel electrophoresis (50 mg/mL kanamycin and 50 mg/mL rifampicin). Positive transformants were grown in 10 mL of LB medium with shaking at 28°C for 24 h. Subsequently, the cells were pelleted by centrifugation at 6000 rpm for 5 min. The cell pellet was resuspended in *Agrobacterium tumefaciens* induction medium (10 mmol/L MES, 10 mmol/L $MgCl_2$, 100 μmol/L acetosyringone, pH 5.8) and incubated at 25°C for 1 h. After measuring the cell suspension concentration at OD_{600}, the bacterial suspension was infiltrated into 4-week-old *N. benthamiana* leaves using a needleless syringe. Three days later, 50 μL of squalene (0.098% in PBS) was injected into the infiltrated leaf areas. After 1 day, the substrate-infiltrated leaves were collected, and 0.2 g of fresh tissue was homogenised in 1 mL methanol using a tissue homogeniser at 25 Hz for 2 min, followed by ultrasonication for 30 min and centrifugation at 12 000 rpm for 10 min. The supernatant was subjected to LC-qTOF-MS analysis. *N. benthamiana* leaves carrying the empty vector were used as the negative control. Each experiment included at least three biological replicates.

Functional expression of candidate CYP450s in *N. benthamiana* was analysed by LC-qTOF-MS using an Agilent 1290 Infinity II UPLC system coupled to a 6546 qTOF mass spectrometer (Agilent). Samples of 5 μL were injected onto an InfinityLab Poroshell 120 EC-C18 column (Agilent, 100 × 2.1 mm, 1.9 μm). The column temperature was held at 42°C, and the flow rate was held constant at 0.35 mL min^{-1}. A gradient using 0.1% v/v formic acid–water as Buffer A and acetonitrile as Buffer B was run as follows: 50%–90% Buffer B from 0 to 3 min, 90% Buffer B from 3 to 8 min, 90%–50% Buffer B from 8 to 9 min, 50% Buffer B from 9 to 11 min. The mass range was 50–1700 m/z. MS acquisition was performed using an ESI ion source operating in positive ion mode, and the data were processed with MassHunter 10.0 software.

4.14 | Heterologous Expression in *Saccharomyces cerevisiae*

We constructed the pCEV-Leu-AtCPR1 vector carrying *A. thaliana* cytochrome P450 reductase (AtCPR1). The pCEV-Leu-AtCPR1 vector, along with pCEV-Trp vectors separately encoding *GpCYP88AB3* and *SynPPDS*, was introduced into the dammarenediol-producing *S. cerevisiae* chassis strain BY-DS (previously developed by our laboratory). The SynPPDS was a codon-optimised version of *Panax notoginseng* PPDS (dammarenediol-II synthase) designed for enhanced heterologous expression in *S. cerevisiae*. Transformants were plated on SD-T-L solid medium, and single colonies were isolated for PCR verification through a 15 min boiling step in 200 μL of 20 mM NaOH at 99°C. Positive clones were streaked on fresh SD-T-L plates and cultured for 3 days. Subsequently, cultures were scaled up, and after 5 days of growth, 6 mL of cell suspension was harvested. For metabolite extraction, an equal volume of glass beads and 1 mL of 70% methanol were added, followed by mechanical disruption using a tissue lyser for 10 min. After centrifugation, the supernatant was collected for LC-qTOF-MS analysis.

4.15 | VIGS Assay of *GpCYP88AB3*

To investigate the function of the *GpCYP88AB3*, VIGS was performed. A gene-specific fragment of *GpCYP88AB3* was cloned into the pTRV2 vector and transformed into the strain GV3101. To infect germinating *G. pentaphyllum* seeds, we employed an in planta transformation system with a 1:1 mixture of *Agrobacterium* cultures containing pTRV1 and pTRV2-*GpCYP88AB3* (Wonbio Biotechnology, Shanghai, China). Seedlings infected with a mixture of pTRV1 and the empty pTRV2 vector served as the negative control, whereas those inoculated with pTRV1 and pTRV2-GpPDS (phytoene desaturase) constituted the positive visual control. After 4 weeks of growth under controlled greenhouse conditions, the plants were harvested for subsequent analysis. The silencing efficiency of *GpCYP88AB3* was confirmed by qRT-PCR. The primers were shown in the Table S9. Additionally, the content of related metabolites was quantified using GC–MS and LC–MS. All experiments were performed with at least three biological replicates, and data were analysed using Student's *t*-test.

After the plants are lyophilised and ground, 20 mg of the powder is taken. It is then extracted with 1 mL of methanol using ultrasonic treatment for 30 min. The mixture is centrifuged at 12 000 rpm for 5 min, and the supernatant is collected for GC–MS analysis. The collected samples were analysed using gas chromatography–mass spectrometry (GC–MS) 7890/5975 C (Agilent Technologies, Santa Clara, CA, USA) equipped with a DB-5 ms fused silica column (Agilent122-5562UI, 60 m × 250 μm × 0.25 μm) with helium as the carrier gas (1 mL/min). The injection volume used was 1 μL. The injection temperature was 250°C, and the solvent delayed for 2 min. The programmed temperature profile was as follows: 160°C held for 1 min, increased at a rate of 20°C/min to 300°C, held for 10 min. The EI ion source was 70 eV electron energy, and the detector temperature of the ion source was 230°C, with SIM mode ion pair 109, 207 and 426 m/z. ChemStation software (Agilent Technologies) was used for the data acquisition and processing. All the aimed compounds were identified by standards.

Author Contributions

P.Z., X.L. and J.-Y.X. conceived the project and designed the study. S.-J.L. and L.H. conducted experiments and acquired and analysed the data. Y.W., W.D., X.J. and Y.Z. helped with data analysis. J.G., H.W. and L.W. helped with LC–MS analysis. P.Z. initiated the study. P.Z. and X.L. wrote the manuscript. P.L., J.-Y.X. and X.L. revised the manuscript. All authors read and agreed to the final article.

Acknowledgements

We thank Open and sharing platform of State Key Laboratory of Natural Medicines, China Pharmaceutical University for kindly providing the LC-QTOF MS (SCIEX X500B).

Funding

This work was funded by the National Natural Science Foundation for Young Scientists of China (grant no. 82404814), the National Natural Science Foundation of China (grant nos. 82574527, 81973414, and 32570265), the Key project at central government level: The ability establishment of sustainable use for valuable Chinese medicine resources

(2060302), 'Double First-Class' University project (CPU2022QZ27), and the Basic Research Program of Jiangsu (BK20252062).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The raw data generated in this study were available in the National Genomics Data Center (BioProject: PRJCA046674).

References

An, C., D. Li, L. Lu, et al. 2024. "Insights Into the Genomic Background of Nine Common Chinese Medicinal Plants by Flow Cytometry and Genome Survey." *Plants (Basel)* 13: 3536. https://doi.org/10.3390/plants13243536.

Chen, H., M. Guo, S. Dong, et al. 2023. "A Chromosome-Scale Genome Assembly of Artemisia Argyi Reveals Unbiased Subgenome Evolution and Key Contributions of Gene Duplication to Volatile Terpenoid Diversity." *Plant Communications* 4: 100516. https://doi.org/10.1016/j.xplc.2023.100516.

Chen, X., C. Tong, X. Zhang, et al. 2021. "A High-Quality *brassica napus* Genome Reveals Expansion of Transposable Elements, Subgenome Evolution and Disease Resistance." *Plant Biotechnology Journal* 19: 615–630. https://doi.org/10.1111/pbi.13493.

Chen, X., C. Yao, J. Hou, et al. 2023. "Systematical Characterization of Gypenosides in Gynostemma Pentaphyllum and the Chemical Composition Variation of Different Origins." *Journal of Pharmaceutical and Biomedical Analysis* 232: 115328. https://doi.org/10.1016/j.jpba.2023.115328.

Cun, Z., J. Zhang, J. Hong, et al. 2024. "Integrated Metabolome and Transcriptome Analysis Reveals the Regulatory Mechanism of Low Nitrogen-Driven Biosynthesis of Saponins and Flavonoids in Panax Notoginseng." *Gene* 901: 148163. https://doi.org/10.1016/j.gene.2024.148163.

David, K. T., J. G. Schraiber, J. G. Crandall, et al. 2025. "Convergent Expansions of Keystone Gene Families Drive Metabolic Innovation in Saccharomycotina Yeasts." *Proceedings of the National Academy of Sciences of the United States of America* 122: e1794802174. https://doi.org/10.1073/pnas.2500165122.

Edger, P. P., T. J. Poorten, R. Vanburen, et al. 2019. "Origin and Evolution of the Octoploid Strawberry Genome." *Nature Genetics* 51: 541–547. https://doi.org/10.1038/s41588-019-0356-4.

Feldman, M., and A. A. Levy. 2012. "Genome Evolution due to Allopolyploidization in Wheat." *Genetics* 192: 763–774. https://doi.org/10.1534/genetics.112.146316.

Gou, S., B. Liu, X. Han, et al. 2018. "Anti-Atherosclerotic Effect of Fermentum Rubrum and Gynostemma Pentaphyllum Mixture in High-Fat Emulsion- and Vitamin D(3)-Induced Atherosclerotic Rats." *Journal of the Chinese Medical Association* 81: 398–408. https://doi.org/10.1016/j.jcma.2017.08.018.

Guo, J., W. Xu, Y. Hu, et al. 2020. "Phylotranscriptomics in Cucurbitaceae Reveal Multiple Whole-Genome Duplications and Key Morphological and Molecular Innovations." *Molecular Plant* 13: 1117–1133. https://doi.org/10.1016/j.molp.2020.05.011.

Han, J., H. Kim, Y. Kwon, and Y. Choi. 2011. "The Cyt p450 Enzyme cyp716a47 Catalyzes the Formation of Protopanaxadiol From Dammarenediol-Ii During Ginsenoside Biosynthesis in *panax ginseng*." *Plant & Cell Physiology* 52: 2062–2073. https://doi.org/10.1093/pcp/pcr150.

Huang, D., J. Li, J. Chen, et al. 2024. "Genome-Wide Identification and Characterization of the Jaz Gene Family in Gynostemma Pentaphyllum Reveals the coi1/Jaz/myc2 Complex Potential Involved in the Regulation of the Meja-Induced Gypenoside Biosynthesis." *Plant Physiology and Biochemistry* 214: 108952. https://doi.org/10.1016/j.plaphy.2024.108952.

Huang, D., R. Ming, S. Xu, et al. 2021. "Chromosome-Level Genome Assembly of Gynostemma Pentaphyllum Provides Insights Into Gypenoside Biosynthesis." *DNA Research* 28: dsab018. https://doi.org/10.1093/dnares/dsab018.

Huang, S., Y. Zhang, X. Wei, et al. 2025. "Chromosome-Level Genome Assembly of an Important Ethnic Medicinal Plant *Callicarpa nudiflora*." *Scientific Data* 12, no. 1: 655. https://doi.org/10.1038/s41597-025-04999-6.

Huang, Y., Y. Wang, Y. Liu, et al. 2022. "Chemical Characterization and Atherosclerosis Alleviation Effects of Gypenosides From Gynostemma Pentaphyllum Through Ameliorating Endothelial Dysfunction via the pcsk9/Lox-1 Pathway." *Journal of Agricultural and Food Chemistry* 70: 11944–11957. https://doi.org/10.1021/acs.jafc.2c02681.

Kang, M., H. Wu, Q. Yang, et al. 2020. "A Chromosome-Scale Genome Assembly of *Isatis indigotica*, An Important Medicinal Plant Used in Traditional Chinese Medicine: An Isatis Genome." *Horticulture Research* 7: 18. https://doi.org/10.1038/s41438-020-0240-5.

Le, D. D., W. Kim, S. Lim, S. C. Kim, and G. Choi. 2021. "Identification of Three Groups of Ginsenoside Biosynthetic Udp-Glycosyltransferases From Gynostemma Pentaphyllum." *Plant Science* 313: 111069. https://doi.org/10.1016/j.plantsci.2021.111069.

Lee, H. S., S. Lim, J. I. Jung, et al. 2019. "Gynostemma Pentaphyllum Extract Ameliorates High-Fat Diet-Induced Obesity in c57bl/6n Mice by Upregulating sirt1." *Nutrients* 11: 2475. https://doi.org/10.3390/nu11102475.

Li, R., J. Wang, Y. Han, and Z. Dai. 2024. "Compound Danshen Yeast 1.0." *Science of Traditional Chinese Medicine* 2: 303–311. https://doi.org/10.1097/st9.0000000000000053.

Li, T., M. Wang, H. Ren, et al. 2025. "Methylation of Isopentenyl Pyrophosphate by C-Methyltransferase and Optimizing Irregular β-Elemene Yield Using UniKP, an Enzyme Kinetic Parameter Prediction Tool." *Journal of Integrative Plant Biology*. https://doi.org/10.1111/jipb.70048.

Liang, H., M. Wang, G. Yang, et al. 2025. "Untargeted Qualitative and Targeted Quantitative Analysis of Saponins Reveal Differential Chemotypes of Gynostemma Pentaphyllum and g. Longipes From Different Geographical Origins." *Food Chemistry* 468: 142412. https://doi.org/10.1016/j.foodchem.2024.142412.

Liang, T., L. Zou, S. Sun, et al. 2019. "Hybrid Sequencing of the Gynostemma Pentaphyllum Transcriptome Provides New Insights Into Gypenoside Biosynthesis." *BMC Genomics* 20: 632. https://doi.org/10.1186/s12864-019-6000-y.

Liu, S., Z. Liu, B. Shao, et al. 2025. "Deciphering the Biosynthetic Pathway of Triterpene Saponins in *Prunella vulgaris*." *Plant Journal* 121: e17220. https://doi.org/10.1111/tpj.17220.

Lou, Y., X. Zheng, Y. Huang, et al. 2021. "New Dammarane-Type Triterpenoid Saponins From Gynostemma Pentaphyllum and Their sirt1 Agonist Activity." *Bioorganic Chemistry* 116: 105357. https://doi.org/10.1016/j.bioorg.2021.105357.

Megalli, S., N. M. Davies, and B. D. Roufogalis. 2006. "Anti-Hyperlipidemic and Hypoglycemic Effects of Gynostemma Pentaphyllum in the Zucker Fatty Rat." *Journal of Pharmacy & Pharmaceutical Sciences* 9: 281–291.

Nguyen, N., T. K. Q. Ha, J. Yang, H. T. T. Pham, and W. K. Oh. 2021. "Triterpenoids From the Genus Gynostemma: Chemistry and Pharmacological Activities." *Journal of Ethnopharmacology* 268: 113574. https://doi.org/10.1016/j.jep.2020.113574.

Niu, Z., F. Zhu, Y. Fan, et al. 2021. "The Chromosome-Level Reference Genome Assembly for *Dendrobium officinale* and its Utility of

Functional Genomics Research and Molecular Breeding Study." *Acta Pharmaceutica Sinica B* 11, no. 7: 2080–2092. https://doi.org/10.1016/j.apsb.2021.01.019.

Paterson, A. H., and C. Queitsch. 2024. "Genome Organization and Botanical Diversity." *Plant Cell* 36: 1186–1204. https://doi.org/10.1093/plcell/koae045.

Su, C., N. Li, R. Ren, et al. 2021. "Progress in the Medicinal Value, Bioactive Compounds, and Pharmacological Activities of Gynostemma Pentaphyllum." *Molecules* 26: 6249. https://doi.org/10.3390/molecules26206249.

Wang, J., P. Sun, Y. Li, et al. 2018. "An Overlooked Paleotetraploidization in Cucurbitaceae." *Molecular Biology and Evolution* 35: 16–26. https://doi.org/10.1093/molbev/msx242.

Wang, L., S. Zhao, H. Cao, and Y. Sun. 2014. "The Isolation and Characterization of Dammarenediol Synthase Gene From Panax Quinquefolius and Its Heterologous Co-Expression With Cytochrome p450 Gene pqd12h in Yeast." *Functional & Integrative Genomics* 14: 545–557. https://doi.org/10.1007/s10142-014-0384-1.

Wang, X., D. Li, X. Guo, et al. 2020. "Comms(n)db-An Automatable Strategy to Identify Compounds From MS Data Sets (Identification of Gypenosides as an Example)." *Journal of Agricultural and Food Chemistry* 68: 11368–11388. https://doi.org/10.1021/acs.jafc.0c03693.

Xue, J., Z. Li, S. Hu, et al. 2023. "The Saururus Chinensis Genome Provides Insights Into the Evolution of Pollination Strategies and Herbaceousness in Magnoliids." *Plant Journal* 113: 1021–1034. https://doi.org/10.1111/tpj.16097.

Yin, X., Y. Xiang, F. Q. Huang, et al. 2023. "Comparative Genomics of the Medicinal Plants *Lonicera macranthoides* and *L. japonica* Provides Insight Into Genus Genome Evolution and Hederagenin-Based Saponin Biosynthesis." *Plant Biotechnology Journal* 21, no. 11: 2209–2223. https://doi.org/10.1111/pbi.14123.

Yuan, X., R. Li, W. He, et al. 2024. "Progress in Identification of UDP-Glycosyltransferases for Ginsenoside Biosynthesis." *Journal of Natural Products* 87, no. 4: 1246–1267. https://doi.org/10.1021/acs.jnatprod.3c00630.

Yun, L., C. Zhang, T. Liang, et al. 2024. "Insights Into Dammarane-Type Triterpenoid Saponin Biosynthesis From the Telomere-To-Telomere Genome of Gynostemma Pentaphyllum." *Plant Communications* 5: 100932. https://doi.org/10.1016/j.xplc.2024.100932.

Zhang, C., M. Rabiee, E. Sayyari, and S. Mirarab. 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, L., Y. Lin, H. Guan, L. Hu, and G. Pan. 2015. "Simultaneous Determination of Gypenoside lvi, Gypenoside Xlvi, 2alpha-Oh-Protopanaxadiol and Their Two Metabolites in Rat Plasma by Lc-Ms/Ms and Its Application to Pharmacokinetic Studies." *Journal of Chromatography. B, Analytical Technologies in the Biomedical and Life Sciences* 1005: 9–16. https://doi.org/10.1016/j.jchromb.2015.09.028.

Zhang, X., Y. Zhao, Y. Kou, et al. 2023. "Diploid Chromosome-Level Reference Genome and Population Genomic Analyses Provide Insights Into Gypenoside Biosynthesis and Demographic Evolution of Gynostemma Pentaphyllum (Cucurbitaceae)." *Horticultural Research* 10: uhac231. https://doi.org/10.1093/hr/uhac231.

Zu, M., Y. Duan, J. Xie, et al. 2021. "Gypenoside Li Arrests the Cell Cycle of Breast Cancer in g0/g1 Phase by Down-Regulating e2f1." *Journal of Ethnopharmacology* 273: 114017. https://doi.org/10.1016/j.jep.2021.114017.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Figure S1:** pbi70598-sup-0001-Figures.pdf. **Table S1:** pbi70598-sup-0002-Tables.xlsx.