frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 12 October 2023
DOI 10.3389/fpls.2023.1273648

Check for updates

OPEN ACCESS

EDITED BY
Chengjiang Ruan,
Dalian Nationalities University, China

REVIEWED BY
Hui Xu,
Guangzhou University of Chinese Medicine, China
Yifan Jiang,
Nanjing Agricultural University, China

*CORRESPONDENCE
Jian-Feng Mao
jianfeng.mao@umu.se
Ilga M. Porth
ilga.porth@sbf.ulaval.ca

RECEIVED 07 August 2023
ACCEPTED 15 September 2023
PUBLISHED 12 October 2023

CITATION
Yan X-M, Zhou S-S, Liu H, Zhao S-W, Tian X-C, Shi T-L, Bao Y-T, Li Z-C, Jia K-H, Nie S, Guo J-F, Kong L, Porth IM and Mao J-F (2023) Unraveling the evolutionary dynamics of the *TPS* gene family in land plants. *Front. Plant Sci.* 14:1273648. doi: 10.3389/fpls.2023.1273648

Unraveling the evolutionary dynamics of the *TPS* gene family in land plants

Xue-Mei Yan[1], Shan-Shan Zhou[1,2], Hui Liu[1], Shi-Wei Zhao[1], Xue-Chan Tian[1], Tian-Le Shi[1], Yu-Tao Bao[1], Zhi-Chao Li[1], Kai-Hua Jia[1,3], Shuai Nie[1,4], Jing-Fang Guo[1,5], Lei Kong[1,6], Ilga M. Porth[7]* and Jian-Feng Mao[1,8]*

[1]National Engineering Research Center of Tree Breeding and Ecological Restoration, State Key Laboratory of Tree Genetics and Breeding, Key Laboratory of Genetics and Breeding in Forest Trees and Ornamental Plants, Ministry of Education, College of Biological Sciences and Technology, Beijing Forestry University, Beijing, China, [2]Shuangyushu No.1 Primary School, Beijing, China, [3]Key Laboratory of Crop Genetic Improvement & Ecology and Physiology, Institute of Crop Germplasm Resources, Shandong Academy of Agricultural Sciences, Jinan, China, [4]Rice Research Institute, Guangdong Academy of Agricultural Sciences & Key Laboratory of Genetics and Breeding of High Quality Rice in Southern China (Co-construction by Ministry and Province), Ministry of Agriculture and Rural Affairs & Guangdong Key Laboratory of New Technology in Rice Breeding, Guangzhou, China, [5]Department of Horticulture and Food, Guangdong Eco-Engineering Polytechnic, Guangzhou, China, [6]Personnel Section, Qufu Nishan National Forest Park Management Service Center, Qufu, China, [7]Département des Sciences du Bois et de la Forêt, Faculté de Foresterie, de Géographie et Géomatique, Université Laval Québec, Québec, QC, Canada, [8]Department of Plant Physiology, Umeå Plant Science Centre, Umeå University, Umeå, Sweden

Terpenes and terpenoids are key natural compounds for plant defense, development, and composition of plant oil. The synthesis and accumulation of a myriad of volatile terpenoid compounds in these plants may dramatically alter the quality and flavor of the oils, which provide great commercial utilization value for oil-producing plants. Terpene synthases (*TPSs*) are important enzymes responsible for terpenic diversity. Investigating the differentiation of the *TPS* gene family could provide valuable theoretical support for the genetic improvement of oil-producing plants. While the origin and function of *TPS* genes have been extensively studied, the exact origin of the initial gene fusion event - it occurred in plants or microbes - remains uncertain. Furthermore, a comprehensive exploration of the *TPS* gene differentiation is still pending. Here, phylogenetic analysis revealed that the fusion of the *TPS* gene likely occurred in the ancestor of land plants, following the acquisition of individual C- and N-terminal domains. Potential mutual transfer of *TPS* genes was observed among microbes and plants. Gene synteny analysis disclosed a differential divergence pattern between TPS-c and TPS-e/f subfamilies involved in primary metabolism and those (TPS-a/b/d/g/h subfamilies) crucial for secondary metabolites. Biosynthetic gene clusters (BGCs) analysis suggested a correlation between lineage divergence and potential natural selection in structuring terpene diversities. This study provides fresh perspectives on the origin and evolution of the *TPS* gene family.

KEYWORDS

terpene synthases, horizontal gene transfer, gene fusion, synteny network, biosynthetic gene clusters

Introduction

Terpenes and terpenoids, constitute the largest and most structurally diverse group of natural compounds in plants and play crucial roles in various physiological and biochemical processes (Bohlmann et al., 1998; Yang et al., 2020; Yu et al., 2020; Zhou and Pichersky, 2020). These plant-derived terpenes offer a wide array of commercially and industrially viable renewable resources for the production of fragrances, flavors, essential oils, and medicinal properties (McGarvey and Croteau, 1995; Chaw et al., 2019). Plant volatile terpenoids (PVTs), largely consisting of monoterpenes and to a lesser extent the sesquiterpenes, are found in high concentrations in oil-producing plants (Little and Croteau, 1999; Li et al., 2021). This great concentration and diversity of PVTs gives plant oils varied properties and flavors, which has led to the extensive utilization of oil-producing plants.

In fact, the PVTs constitute the major, and often the characteristic, components of essential oils (Little and Croteau, 1999; Li et al., 2021). *Eucalyptus* leaves are enriched in volatile terpenes such as 1-8 cineole, α-terpinene, γ-terpinene, carvacrol, thymol and limonene (King et al., 2006; Naidoo et al., 2014; Kainat et al., 2019). Citronellol, geraniol, and nerol are the three major components of rose oil (Li et al., 2021). The mint family, Lamiaceae, produces mint oils that are rich in monoterpenes and sesquiterpenes (Vining et al., 2017). Tea tree oil, another terpene-rich essential oil, is commonly used in cosmetics and skin care products due to its antimicrobial properties (Carson et al., 2006). Significant cases of terpene richness can also be found in conifer leaves and wood, which contain abundant monoterpenes and diterpenoids (Trapp and Croteau, 2001), as well as in yellowhorn (*Xanthoceras sorbifolium*), an underutilized oil-producing tree that yields triterpenoids (Wang et al., 2017; Chen et al., 2020).

The enormous diversity of plant terpenes in specialized metabolism can be primarily attributed to the diverse terpene skeletons generated by typical plant terpene synthases (*TPSs*) (Tholl, 2015; Alquézar et al., 2017; Pichersky and Raguso, 2018). Through studying the polymorphism of *TPS* genes, genetic improvement can be carried out to breed oil-producing plant varieties with high volatile terpenoid compound content. The *TPSs* belong to a mid-sized gene family and are prevalent in land plants (Chen et al., 2011; Jia et al., 2022). Each full-length plant *TPS* contains two conserved domains identified by their Pfam models: PF03936 (C-terminal) and PF01397 (N-terminal) (Chen et al., 2021). Based on the reaction mechanism and the products formed, *TPS* enzymes can be classified into two classes. Class I enzymes active sites, which are located in the C-terminal region, are characterized by highly conserved aspartate-rich DDxxD and "NSE/DTE" motifs found within an "α-domain". In contrast, the active site of Class II enzymes is located in the N-terminal region between a pair of alphahelical double-barrel domains, known as the "β-domain" and an additional "γ-domain", this site utilizes a "DxDD" motif (Yang et al., 2020; Zhou and Pichersky, 2020; Jia et al., 2022). In some non-seed land plants, *ent*-copalyl diphosphate/*ent*-kaurene synthase (*CPS/KS*) functions as a bifunctional "$\alpha\beta\gamma$-tridomain" enzyme with both class I and class II activity at the N-terminal and C-terminal, respectively (Hayashi et al., 2006; Jia et al., 2018; Yang et al., 2020). This gene can synthesize both *CPS* and *KS*, which are the two key di-*TPSs* that sequentially catalyze the biosynthetic intermediate *ent*-kaurene of gibberellic acids (GAs) (Morrone et al., 2009; Kawaide et al., 2011). However, seed plants possess separate genes for *CPS* and *KS* synthesis but still contain the ancestral "$\alpha\beta\gamma$-tridomain" architecture (Köksal et al., 2011; Zi et al., 2014; Jia et al., 2022). Consequently, the bifunctional "$\alpha\beta\gamma$-tridomain" *TPS* is widely believed to be an ancestral gene.

Based on the structural and functional similarity between plant diterpene cyclases and bacterial diterpene cyclases, it is believed that the ancestral land plant "$\alpha\beta\gamma$-tridomain" (*CPS/KC*) gene evolved through the fusion of the "$\beta\gamma$-didomain" (*CPS*) and the "α-domain" (*KS*). Subsequent loss of the "γ-domain" in the "$\alpha\beta\gamma$-tridomain" led to the origin of the "$\alpha\beta$-didomain" genes (Cao et al., 2010; Köksal et al., 2011; Smanski et al., 2012). However, it remains unclear whether the fusion of the "$\beta\gamma$-didomain" (*CPS*) and the "α-domain" (*KS*) occurred in the ancestral land plants after acquiring them through horizontal gene transfer (HGT) from bacteria, or if the fusion first took place in microbes and was then acquired by ancestral land plants via HGT (Jia et al., 2022). Thus, phylogenetic analysis with comprehensive sampling of terpene synthase sequences from plants and microbes is necessary.

The early evolution of the plant *TPS* family has been reported with substantial evidence. It is likely that the ancestral bifunctional *TPS* gene underwent at least two duplications, giving rise to three ancient *TPS* lineages that led to the present subfamilies TPS-c, TPS-e/f, and TPS-h/d/a/b/g (Jia et al., 2022). The separate *CPS* genes (class II), along with the extant *CPS/KS* genes, form the TPS-c subfamily; the separate *KS* genes (class I) gave rise to the TPS-e/f subfamily. Meanwhile, the angiosperm-specific TPS-a/b/g, gymnosperm-specific TPS-d, and TPS-h subfamilies are dedicated to secondary metabolism and synthesize mono-, sesqui-, diterpenes, among others (Yang et al., 2020; Jia et al., 2022). The conservation and differentiation dynamics of *TPS* subfamilies involved in primary and secondary metabolism across plant lineages are still opaque. Examining the gene synteny of *TPS* genes across various species could provide crucial information to answer fundamental questions about the evolution of this important gene family (Zhao et al., 2017).

In eukaryotic organisms, nonhomologous genes that jointly encode the biosynthetic enzymes in a specialized metabolic pathway are often co-localized within the genome. These local clusters of genes are referred to as biosynthetic gene clusters (BGCs) (Kautsar et al., 2017; Nützmann et al., 2018; Rokas et al., 2018; Witjes et al., 2019). In plants, species-specific BGCs are formed through processes such as gene duplications, neofunctionalizations, subfunctionalizations, and relocations. These functional units are inherited and provide a selection advantage in response to various biotic stresses (Qi et al., 2004; Zhou et al., 2016; Wu et al., 2022). Some studies on terpene-related BGCs demonstrated that minor differences in the structures of the pathway end-products leading to diversification have arisen from gene duplication, random mutations, and neofunctionalization (Zhou et al., 2016; Nützmann et al., 2018; Rokas et al., 2018; Liu et al., 2020). Comparing a candidate BGC with homologous genomic loci across multiple plants can provide important information about its evolutionary conservation or diversification (Kautsar et al., 2017). Although BGC studies have recently received increasing attention and have been conducted in some plants, interspecific terpene-related BGC homology analysis is still lacking.

Here, we conducted a comprehensive investigation into the origin and evolutionary dynamics of plant *TPS* genes by utilizing a broad phylogenetic sampling and gene synteny-based comparative genomic analysis. Our results revealed potential mutual transfer of *TPS* genes between microbes and plants. Additionally, we found it likely that the PF03936 (C-terminal) and PF01397 (N-terminal) domains were individually acquired from microbes through HGT, and their fusion event probably occurred in the ancestor of land plants. Moreover, our comparative genomic analysis uncovered notable patterns, where the subfamilies involved in primary metabolism were conserved, while the subfamilies exclusively associated with secondary metabolite production exhibited significant divergence and radiation. We also observed a substantial expansion of *TPS* genes and the insertion or deletion of other metabolic genes in both homologous and species-specific BGCs. These findings enhance our understanding of plant terpenes diversity from multiple perspectives and provide insights relevant to studying the evolution of *TPS* gene families. Furthermore, our results offer valuable references for breeding oil-producing plants that contain high concentrations of volatile terpenoids.

Materials and methods

Species selection and gene identification

We analyzed assembled genome sequences and their annotated protein sequences from 74 plant species to identify *TPS* genes. The sampled plant species spanned diverse lineages, including Chlorophyta, Streptophyta, Bryophytes (Liverworts and Mosses), Lycopodiophyta, Gymnospermae (Cupressales, Gnetales, Ginkgoales), and Angiosperm (Apiales, Asterales, basal Angiosperms, Brassicales, Lamiales, Laurales, Malpighiales, Myrtales, Poales, Rosales, Sapindales, Solanales, Vitales) (Supplementary Table 1). We employed the hmmscan v3.2.1 (Mistry et al., 2013) with an E-value of 1e-5 to identify candidate *TPS* genes containing at least one conserved Pfam (Bateman et al., 2004) domain (PF01397 and/or PF03936). Subsequently, we verified the conserved domains via Conserved Domain Database (CDD; https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). Based on the presence of both PF01397 and PF03936 domains, only the PF01397 domain, or only the PF03936 domain, we classified these high-confidence *TPS* homologies into three categories: full-length *TPS* genes, PF01397-domain-genes, and PF03936-domain-genes.

To extensively mine *TPS* sequence among various microbes, we identified putative *TPS* sequences by searching the significantly matched proteins (containing at least one conserved domain PF01397 and/or PF03936) in NR databases using hmmscan v3.2.1 (Mistry et al., 2013) (E-value = 1e-5). The NR database were obtained from https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/ (accessed date: 2022-10-31). We separately extracted all protein sequences of bacteria, fungi and archaea from the database using TaxonKit, a command-line interface tool for handling NCBI taxonomy data (Shen and Ren, 2021). We extended the classification method of plant *TPS* genes to include sequences from bacteria, fungi and archaea. Sequences that matched either the PF01397 or PF03936 domains were classified into full-length *TPS* genes, PF01397-domain-sequences and PF03936-domain-sequences. Furthermore, we determined the species' taxonomic origins of PF01397-domain-sequences and PF03936-domain-sequences using the taxonomizr package (versions 0.9.3, https://www.rdocumentation.org/packages/taxonomizr/versions/0.9.3) in R. The original taxon was listed in a matrix that included six hierarchical levels, i.e., phylum, class, order, family, genus, and species.

Phylogenetic reconstruction

To classify the full-length *TPS* genes from plants, we separately constructed phylogenetic trees for full-length *TPS* genes from 62 angiosperm species and non-angiosperm species, including *T. wallichiana*, *Sequoiadendron giganteum*, *G. montanum*, *Ginkgo biloba*, *M. polymorpha*, *P. patens*, and *S. moellendorffii*. To investigate the origins of two full-length *TPS* genes (*WP_145718914.1* and *WP_260645699.1*) identified in two bacteria, we further performed phylogenetic reconstruction for these two genes in conjunction with 40 plant genes, and conservation motifs of these genes were identified using MEME (https://meme-suite.org/meme/tools/meme) (Bailey et al., 2009). Sequences alignment was prepared using MAFFT v7.407 (Katoh et al., 2002; Katoh and Standley, 2013) with "–anysymbol", and poorly aligned regions were trimmed using trimAl v1.2.rev59 (Capella-Gutiérrez et al., 2009) with "-gt 0.4 -st 0.001".

To explore the origin and relationships of the PF03936 and PF01397 domains in *TPS* genes across plants and microbes, we constructed separate unrooted phylogenetic trees using sequence data of these domains extracted from plants, fungi, archaea and bacteria. We clustered the sequences (size >100 aa) of PF03936 and PF01397 domains with at least 80% identity using CD-HIT with parameters -c 0.80 -n 5 -M 16000 -d 0 -T 8 and selected the longest sequences in each cluster for further analysis. For the PF03936 domain, we constructed a phylogenetic tree using representative sequences from archaea and plants, and randomly sampled 1000 sequences from bacteria and fungi using seqkit (Shen et al., 2016). For the PF01397 domain, we used all representative sequences from bacteria, archaea, fungi, and plants to construct the phylogenetic tree. Sequences alignment was prepared using MAFFT v7.407 (Katoh et al., 2002; Katoh and Standley, 2013) with "–anysymbol" option and without trimming.

We constructed all phylogenetic trees using either alignments or trimmed alignments, utilizing IQ-TREE v2.0.3 (Nguyen et al., 2014) with 1,000 replications of ultrafast bootstrap and Shimodaira-Hasegawa-like approximate likelihood-ratio (SH-aLRT) test. The best-fit models chosen by ModelFinder (Kalyaanamoorthy et al., 2017) were as follows: "JTT+R7" for the angiosperm full-length genes tree, "JTT+F+R6" for the non-angiosperm full-length genes tree, "JTT+F+R3" for the full-length bacterial genes tree, "LG+R10" for the PF03936 domain sequences tree, and "JTT+F+R6" for the PF01397 domain sequences tree. We interpreted and visualized those phylogenetic trees using the online tool iTOL v6 (Letunic and Bork, 2007).

Two full-length *TPS* genes (*WP_145718914.1* and *WP_260645699.1*) from soil bacteria and further confirmation

Two full-length *TPS* genes (*WP_145718914.1* and *WP_260645699.1*) were identified from soil bacteria. To further confirm their occurrence, we conducted a homology analysis of their flanking regions among the conspecific and congeneric genomes. Homologous search was conducted by BLASTn, and significant match was identified for an E-value of less than 1e-5 and identity greater than 70%. The homologous analysis for *WP_145718914.1* (found in VLLG01000006.1 sequence of *C. japonensi*) was conducted among 41 *Chitinophaga* species, and the analysis for *WP_260645699.1* (found in JAOBSP010000025.1 sequence of *S. aureus*) was conducted among 96 *S. aureus* strains and 59 *Staphylococcus* species. The sequence homology of this flanking region analysis was visualized using the command 'python -m jcvi.graphics.synteny' of MCscan (Python version) (Wang et al., 2012).

Synteny network analyses and phylogenetic profiling

We constructed a genomic synteny network of plant *TPS* genes based on the processes designed in the SynNet pipeline (Zhao et al., 2017; Zhao and Schranz, 2019). First, we performed a reciprocal all-against-all BLAST search using Diamond v0.9.22.123 (Buchfink et al., 2015) with -k 5 against sequences in the proteomes of 74 studied plants (Supplementary Table 1). After that, the genomic collinearity between all possible pairwise genome combinations using MCScanX (Tang et al., 2008; Wang et al., 2012) with default parameters (minimum match size for a collinear block = 5 genes, max gaps allowed = 25 genes) were calculated. Subsequently, we extracted all possible syntenic gene pairs of the full-length *TPS* genes, PF03936-domain-genes and PF01397-domain-genes to construct their respective synteny networks.

We further imported the full-length *TPS* gene synteny network into CFinder v.2.0.6 to detect potential *k*-clique communities (subnetworks) with *k*=3 (Derényi et al., 2005; Palla et al., 2005; Fortunato, 2010). The *k*-clique corresponds to the two-by-two connection of *k* nodes to each other (e.g., a *k*-clique of *k*=3 is equivalent to a triangle) (Zhao et al., 2017). After *k*-clique percolation, we visualized all networks using Gephi v0.9.2 (Bastian et al., 2009), considering *TPS* subfamily and plant lineages information for downstream analysis. We clustered the synteny networks of full-length *TPS* genes, PF03936-domain-genes and PF01397-domain-genes using infomap (Rosvall and Bergstrom, 2008), and counted the number of genes in resulting clusters. These derived gene numbers were mapped back to the 74 species (Supplementary Table 1). Angiosperms were arranged according to their phylogenetic treatment developed in APG (Angiosperm Phylogeny Group) IV (The Angiosperm Phylogeny et al., 2016), while the gymnosperms and early land plants arranged according to the previous studies (Cheng et al., 2019; Zhang et al., 2020).

Identification of syntenic terpene-related BGCs among species

We identified BGCs from the whole genomes of our 74 sampled plants (Supplementary Table 1) using antiSMASH v3.0.5 (PlantiSMASH python version) (Kautsar et al., 2017) with the parameters of "–taxon plants –debug –cdh-cutoff 0.5 –min-domain-number". We selected a representative species with high genome quality and assembly completeness as the reference genome for each plant lineage. Specifically, we chose *B. napus* of Brassicales, *Citrus medica* of Sapindales, *E. grandis* of Myrtales, *H. brasiliensis* of Malpighiales, *Pyrus bretschneideri* of Rosales, *V. vinifera* of Vitales, *Helianthus annuus* of Asterales, *O. basilicum* of Lamiales, *Capsicum annuum* of Solanales, *S. cereale* of Poales, *Litsea* cubeba of Laurales, *A. trichopoda* of basal_Angiosperms, *T. wallichiana* of gymnosperms, and *S. moellendorffii*, *P. patens*, *M. polymorpha* for non-seed plants.

We performed pairwise synteny region search for BGCs of these representative species and other 69 land plants using the 'python -m jcvi.compara.synteny mcscan' and 'python -m jcvi.formats.base join' commands with the MCscan (Python version) (Tang et al., 2008) default parameters. Based on the presence of BGCs similar to those in the reference genome on the synteny blocks, we further identified highly conserved BGCs, lineage-specific BGCs, and species-specific BGCs. Microsynteny visualization was prepared using the command 'python -m jcvi.graphics.synteny'.

Results

Distribution and evolution of *TPS* genes in plants

After extensive sequences mining, we found that full-length *TPS* genes, PF03936-domain-genes, and PF01397-domain-genes (see Materials and Methods for definition of gene categories) were unevenly distributed among the 74 plants (69 land and 5 lower plant species) (Supplementary Table 1). We identified a total of 3,600 full-length *TPS* genes (3,167 of which were longer than 350 amino acids) in all land plant species. Additionally, 513 PF01397-domain-genes were discovered exclusively in seed plant species, and 1,049 PF03936-domain-genes in all land plants except for the moss *Physcomitrella patens* (Figures 1A–C; Supplementary Table 2). In the five lower plant species, only five PF03936-domain-genes were found in *Klebsormidium nitens* (Figure 1C; Supplementary Table 2).

To classify the full-length *TPS* genes, we constructed two unrooted *TPS* phylogenetic trees for 3,386 genes from 62 angiosperms and 241 genes from 7 non-angiosperms, respectively

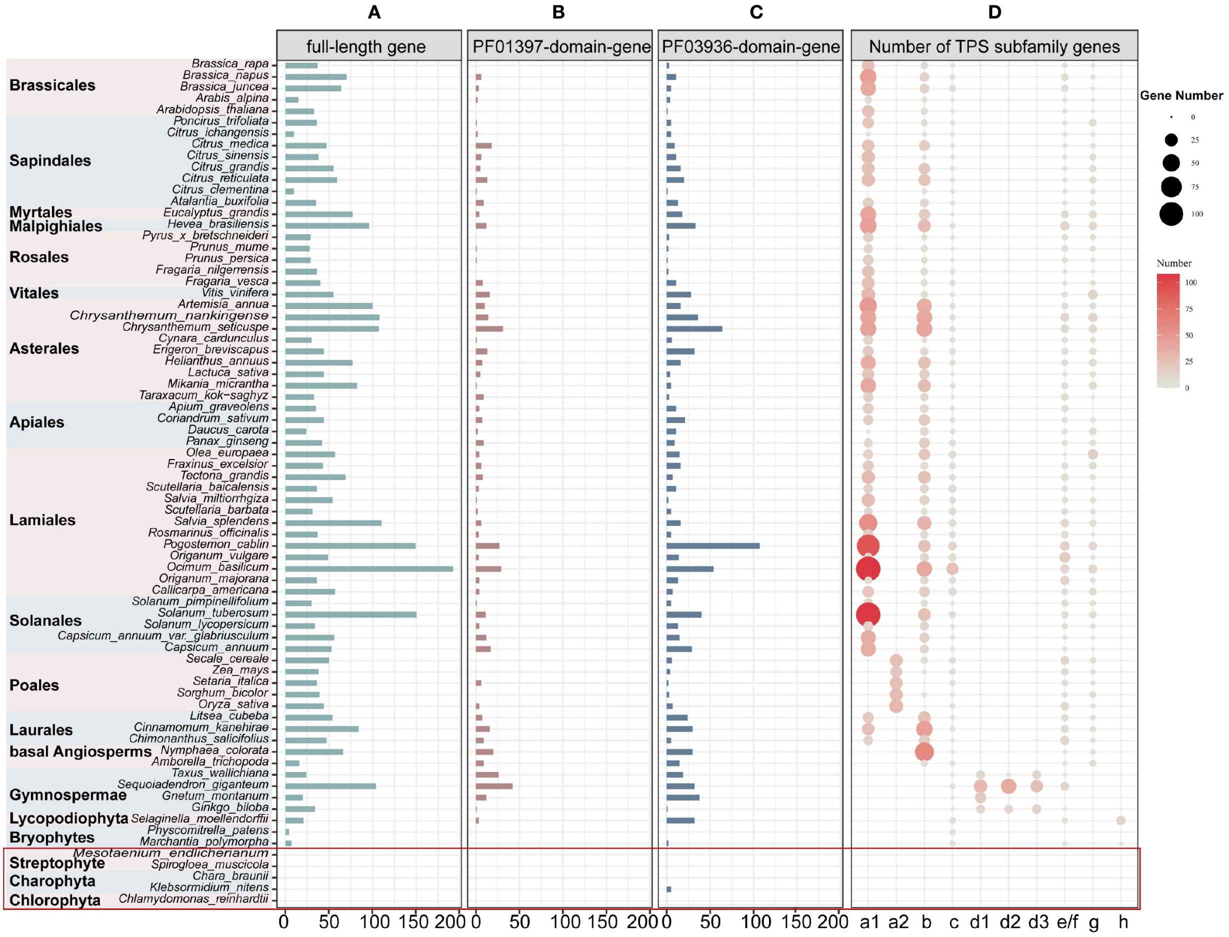

FIGURE 1
Distribution and classifications of *TPS* genes in 74 studied plant species across 19 orders. **(A-C)** Number of *TPS* genes (including full-length, PF01397-domain-genes, and PF03936-domain-genes). The red square highlights lower plant species. The "full-length gene" represents genes with both "PF01397" and "PF03936" functional domains; "PF01397-domain-gene" and "PF03936-domain-gene" represent plant sequences with only one functional domain of either "PF01397" or "PF03936". **(D)** The size and color of the bubbles represent the full-length gene number within each subfamily.

(Supplementary Figures 1, 2). Based on the previously characterized genes (Supplementary Tables 3, 4), these genes were classed into the seven subfamilies TPS-a, TPS-b, TPS-c, TPS-d, TPS-e/f, TPS-g, and TPS-h (Figure 1D; Supplementary Figures 1, 2; Supplementary Table 5). The TPS-c and TPS-e/f subfamilies play a role in primary metabolism and are shared among all angiosperm lineages (Figure 1D; Supplementary Table 6). However, the divergence of these two subfamilies is inconsistent within both non-seed plants and gymnosperms. The first divergence was observed in liverwort, *Marchantia polymorpha*, while the TPS-e/f subfamily was absent in the moss *P. patens* (Figure 1D; Supplementary Table 6). Among gymnosperms, *Gnetum montanum* also lacks the TPS-e/f subfamilies (Figure 1D; Supplementary Table 6). Consistent with previous research (Chen et al., 2011), the TPS-e/f clade in angiosperms underwent a deep divergence, forming two subclades, TPS-f and TPS-e, with TPS-f being dicot-specific (Supplementary Figure 1). The divergence of TPS-c and TPS-e/f subfamilies suggests parallel subfunctionalization within each lineage.

The subfamilies TPS-a/b/d/g/h, specialized in secondary metabolism, exhibited lineage-specific differences. The TPS-a/b/g subfamilies were angiosperm-specific, with TPS-a further divided into dicot-specific TPS-a1 and monocot-specific TPS-a2 clades. However, TPS-a was absent in the basal angiosperm species *Nymphaea colorata* and *Amborella trichopoda* (Figure 1D; Supplementary Table 6), suggesting that TPS-a could represent a novel subfamily that emerged after the divergence of Mesangiospermae from basal angiosperms. TPS-d was gymnosperm-specific and further divided into TPS-d1, TPS-d2, and TPS-d3 clades (Figure 1D; Supplementary Table 6). The clades TPS-d2 and TPS-d3 were not present in *G. montanum* (Figure 1D; Supplementary Table 6), potentially indicating a loss of these two sub-clades in *G. montanum* during its evolution or that the subfunctionalization of the TPS-d subfamily after *G. montanum* diverged from other gymnosperms species. TPS-h was identified in both the liverwort *M. polymorpha* and Lycopodiophyta *Selaginella moellendorffii* (Figure 1D; Supplementary Table 6), suggesting that divergence of the TPS-h clade occurred before the split of liverwort and Lycopodiophyta. In addition, the TPS-b subfamily was absent in monocot species *Oryza sativa* and *Setaria italica*, and the TPS-g subfamily was absent in the dicot species *Apium graveolens* (Figure 1D; Supplementary Table 6).

These absences may have resulted from independent losses of entire the subfamilies in specific plants.

Prevalent distribution of *TPS* PF03936 and PF01397 domain sequences in microbes

The domain sequences of PF01397 and PF03936 were found to be prevalent among various microbial species (Supplementary Tables 7–9). Approximately 45% of bacterial families, 26% of archaeal families, and 20% of fungal families were found to possess sequences with PF01397 domain. Meanwhile, around 58% of bacterial families, 48% of archaeal families, and 33% of fungal families contained sequences with PF03936 domain (Table 1). Among these sequences, 1.35% to 5.56% of those containing the PF01397 domain and 32.84% to 90.70% of those containing the PF03936 domain were found to be significantly related to conserved terpene synthesis domains with an E-value of 1e-5 (Table 1), such as "Isoprenoid_Biosyn_C1 superfamily", "Terpene_syn_C_2", "PLN02279 super family (ent-kaur-16-ene synthase)", "PLN02592 superfamily (ent-copalyl diphosphate synthase)", "SQHop_cyclase_C superfamily", "squalene_cyclas superfamily", and others.

However, full-length *TPS* genes were exceedingly rare in microbes. Only four soil bacterial sequences (*WP_145718914.1*, *WP_260645699.1*, *WP_232298052.1*, *KIO47690.1*) were found to match both the PF03936 and PF01397 domains (Table 1). The *WP_145718914.1* and *WP_260645699.1* genes encode potential proteins longer than 500 amino acids (Table 1). Further phylogenetic and conservation analysis revealed the *WP_260645699.1* gene shared highly conserved motifs with gymnosperm-specific TPS-d1 genes, while the *WP_145718914.1* gene displayed a region of conserved motifs towards the 5' end, shared with multiple subfamilies, and demonstrated larger differentiation at the 3' end (Supplementary Figure 3A).

The flanking sequences of the *WP_145718914.1* gene in *Chitinophaga japonensis* shared homologous sequences with 10 *Chitinophaga* species (Supplementary Figure 3B; Supplementary Tables 10, 11). After remapping the genomic sequencing data, the gene region of the *WP_145718914.1* showed consistent depth of coverage across the gene body, start and end positions and these gene body coverages were comparable to the genome-wide average (Supplementary Figure 3C). This even depth of coverage across the entire gene partially mitigated the impact of genome assembly errors and sequencing contamination. These results supported the presence of *WP_145718914.1*, a plant origin full-length *TPS* genes in a soil bacterium genome, and suggests it may be an intact gene that has been functionally integrated into the genome through horizontal transfer. In contrast, we did not find any homologous fragments of the flanking sequence of the *WP_260645699.1* gene from *Staphylococcus aureus* in the 96 *S. aureus* strains (Supplementary Table 12) and 59 *Staphylococcus* genus species

TABLE 1 Statistics of PF03936 or PF01397 domains sequences derived from bacteria, archaea and fungi in the NR database.

			Bacteria	Archaea	Fungi
NR database		**Sequences number**	397285288	9525243	27330891
		Species number	>115384	>4252	>34345
		Family number	>665	>65	>777
PF03936-domain sequences	**hmmscan**	**Sequences number**	22789	670	8268
		Species number	>7520	>278	>1457
		Family number	>383	>31	>256
	CDD	**Sequence number**	10894	220	7499
PF01397-domain sequences	**hmmscan**	**Sequences number**	7338	234	778
		Species number	>2956	>91	>520
		Family number	>302	>17	>152
	CDD	**Sequence number**	99	13	21
Full-length genes	**Gene IDs**	**Family**	**Species**	**CDD domain (E-value = 1e-5)**	**Sequence length (aa)**
	WP_145718914.1	Chitinophagaceae	*Chitinophaga japonensis*	PLN02592 and Terpene_syn_C2	785
	WP_260645699.1	Staphylococcaceae	*Staphylococcus aureus*	Terpene_cyclase_plant_C1	584
	WP_232298052.1	Nitrosomonadaceae	*Nitrosospira* sp. NpAV	Isoprenoid_Biosyn_C1 superfamily	152
	KIO47690.1	Nitrosomonadaceae	*Nitrosospira* sp.	Isoprenoid_Biosyn_C1 superfamily	177

Detailed information of four full-length genes containing both PF01397 and PF03936 domains found in bacteria are shown. "hmmscan" refers to the number of microbial sequences that closely match the Pfam models of PF01397 and PF03936, as identified by hmmscan with E-value = 1e-5; "CDD" indicates the number of sequences of microbes related to terpene synthesis and verified by CDD (https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi) analysis.

(Supplementary Table 13). Considering that the sequence in which this gene is located is only 2224 bp long, the presence of the *WP_260645699.1* gene in a bacterial genome remains doubtful.

Phylogenetic relationships of PF03936 and PF01397 domain sequences

The phylogenetic relationship of both the PF03936 and PF01397 domain showed a similar topology, indicative of a deep divergence that occurred between plant and microbial (bacteria, archaea and fungi) sequences (Figures 2A, B). Deep divergence was also evident among major microbial clades and between angiosperm-specific TPS-a/b/g subfamilies and four other subfamilies (TPS-c, TS-e/f, TPS-d, TPS-h) (Figures 2A, B). The phylogenetic relationships of the two *TPS* domain sequences in plants were largely consistent with the full-length *TPS* genes (Figures 2A, B; Supplementary Figures 1, 2), suggesting that all full-length *TPS* genes in plants descended from a common ancestor. Additionally, the trees showed that a few plant PF01397/PF03936-domain sequences were located within microbial clades and some microbial sequences were found within plant clades, indicating potential mutual transfer between microbes and plants (Figures 2A, B).

Of particular interest is the phylogenetic position of the soil bacterial genes *WP_145718914.1* and *WP_260645699.1*. Both the

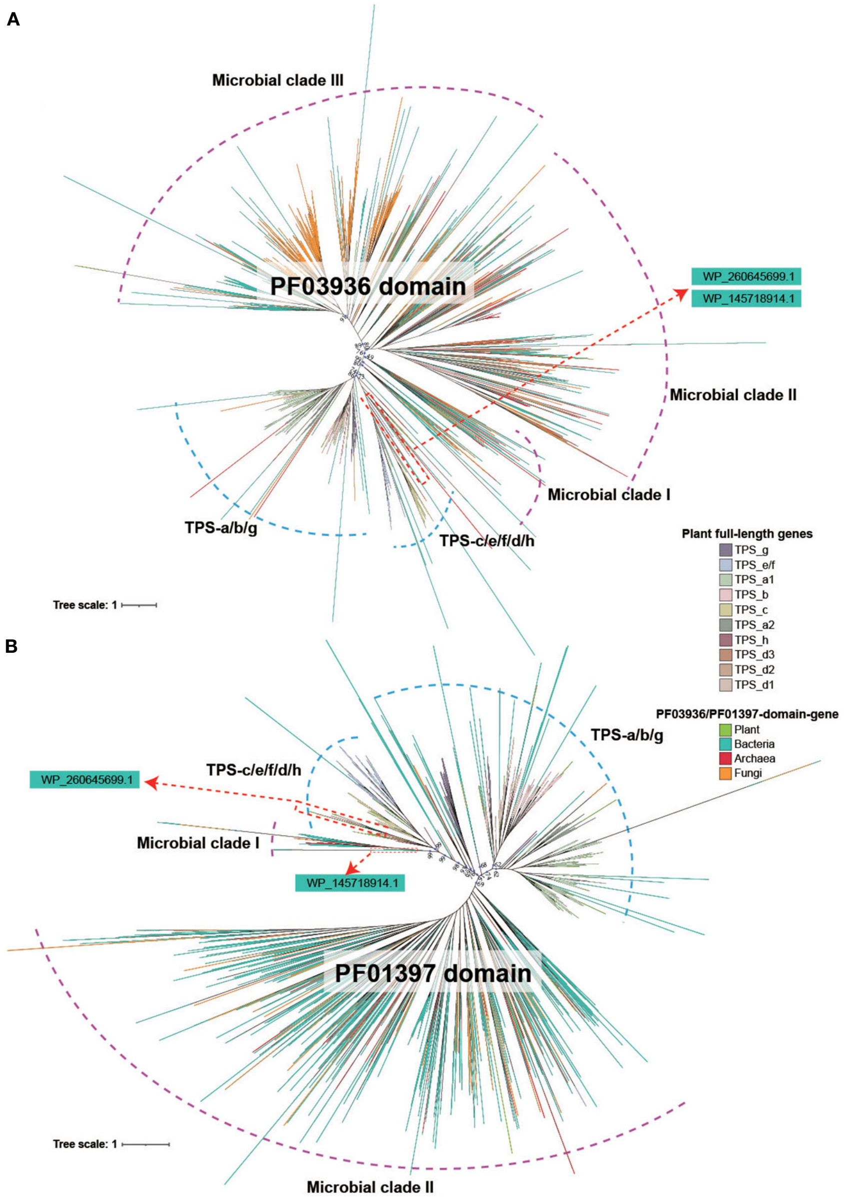

FIGURE 2
Phylogenetic tree of PF03936 and PF01397 domain sequences in plant and microbial. **(A)** Tree reconstructed based on PF03936 domain sequences from plant, bacteria, archaea and fungi. **(B)** Tree reconstructed based on PF01397 domain sequences from plant, bacteria, archaea and fungi. The color of the branch denotes sequence origins from different biological groups or *TPS* subfamilies.

PF01397 and PF03936 domain sequences of *WP_260645699.1* were located in the gymnosperm-specific TPS-d1 subfamily clade (Figures 2A, B). The PF03936 domain of *WP_145718914.1* gene was located in TPS-h clade, while the PF01397 domain was positioned closer to the "Microbial clade I" (Figures 2A, B). This further suggests that the *WP_260645699.1* gene is highly conserved with TPS-d subfamily genes. A significant divergence of PF01397 domain was found between *WP_145718914.1* and the plant genes.

Conserved and lineage-specific divergence of *TPS* genes in land plants

To shed light on the conserved and lineage-specific evolutionary landscape of *TPS* genes in land plants, we further examined the inter- and intra-specific synteny relationships of plant *TPS* genes. Through gene synteny network clustering, we identified a total of 122 synteny clusters of full-length *TPS* genes, 201 synteny clusters of PF03936-domain-genes, and 121 synteny clusters of PF01397-domain-genes. These clusters revealed a clear conserved and lineage-specific pattern of *TPS* genes among angiosperm plants (Figure 3A; Supplementary Figure 4).

Only a small number of gymnosperms TPS-c and TPS-e/f genes were found clustered in lineage-specific clusters in the syntenic network, whereas the TPS-h and gymnosperm-specific TPS-d subfamilies were absent from the synteny network (Figure 3B; Supplementary Figure 5; Supplementary Table 14). This is likely due to extreme phylogenetic distance and sampling biases between angiosperms and non-angiosperm species. In the conserved clusters (where cluster genes are shared among multiple lineages), the genes mainly belong to the TPS-c, TPS-e/f, TPS-b, and TPS-g subfamilies. Additionally, the Lamiales, Solanales, Sapindales, and Brassicales lineage-specific clusters (where cluster genes are primarily shared in only one lineage) mainly consist of TPS-a1 genes, and the Poales lineage-specific clusters primarily contain TPS-a2 genes, and the Asterales lineage-specific clusters mostly consist of TPS-b genes (Figure 3B; Supplementary Table 14). In the Lamiales, which include important spice plants, we found that the lineage-specific

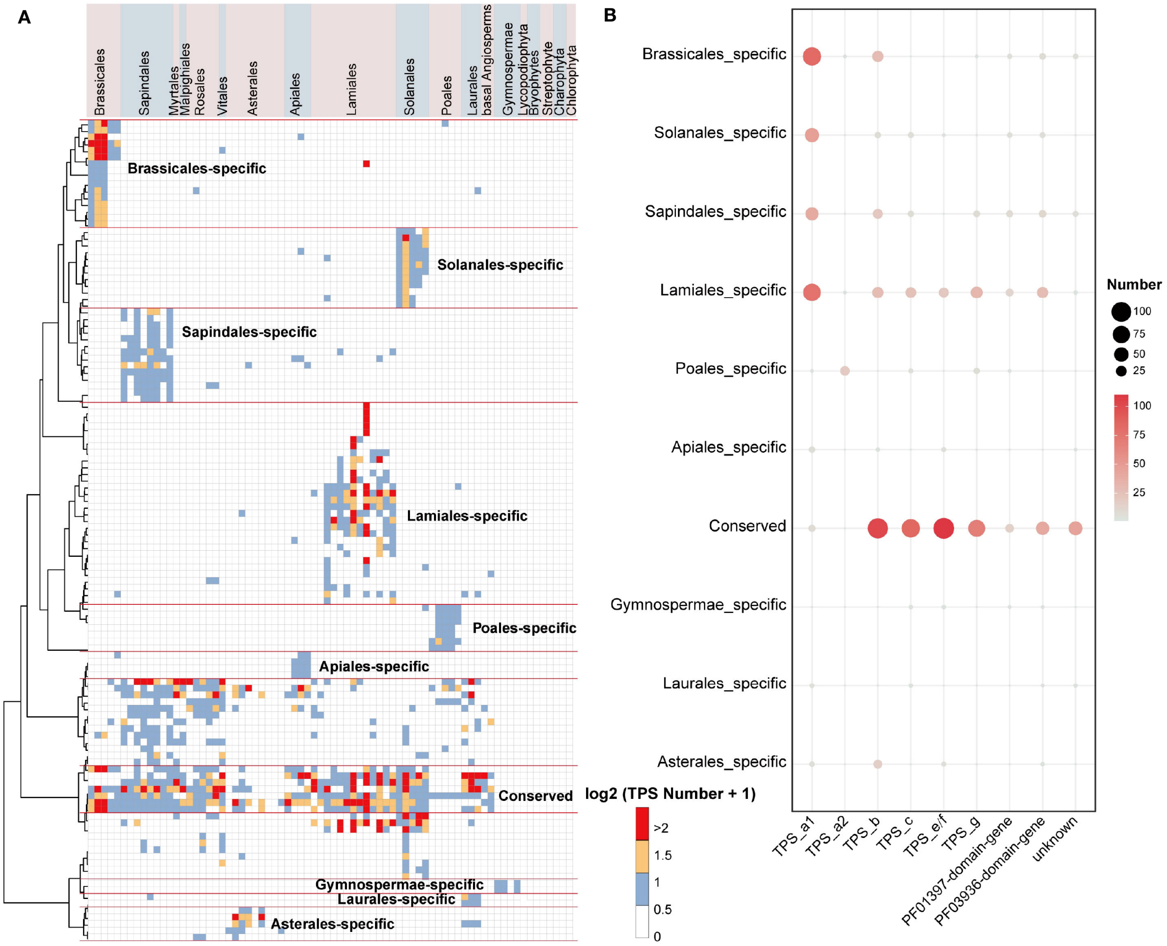

FIGURE 3
Gene synteny clusters of full-length *TPS* genes in 74 plant species across 19 orders. **(A)** Phylogenetic profiling of the 122 network clusters obtained by infomap (Rosvall and Bergstrom, 2008). The color scale within the heatmap represents the number (log2(gene number + 1)) of full-length *TPS* genes found in each cluster for a given species. Red lines distinguish the specific and conserved *TPS* gene clusters in plant lineages. **(B)** Bubble chart illustrating the number of *TPS* genes in the conserved and lineage-specific clusters. "NA" denotes sequences that have not matched the PF01397 or PF03936 domains. The bubble size and color scheme indicate the gene number.

cluster contains a large number of TPS-a subfamily genes along with many TPS-b, TPS-c, TPS-e/f, and TPS-g subfamilies genes (Figures 3A, B). This may be a key indication of the diversity of terpenes in this lineage.

These conserved and lineage-specific clusters were clearly evident in each subfamily subnetwork (Figures 4B–K; Supplementary Figure 5), demonstrating that *TPS* genes have undergone duplication and functional divergence in different plant lineages over evolutionary time. In the TPS-c subfamilies subnetwork, we identified one major cluster contains genes from all the angiosperm lineages, indicating broad conservation across flowering plants. We also found five lineage-specific clusters composed of genes belonging only to Lamiales, Solanales, Sapindales, Laurales, and Gymnospermae, respectively (Figures 4B, D). The TPS-f subfamily is derived from the TPS-e subfamily and has been identified as TPS-e/f (Chen et al., 2011). In our TPS-e/f subnetwork, the previously characterized TPS-e and TPS-f related genes (for gene details, see Supplementary Table 4) were found in two distinct clusters without any synteny relationships between them. This suggests functional divergence of the duplicated genes (Figures 4C, E). Notably, the TPS-e cluster genes were found in all angiosperm lineages, while monocots lacked genes in the TPS-f cluster (Figures 4C, E). This further indicates that TPS-f subfamilies may have evolved independently in dicots after the divergence of monocots and dicots.

We further examined pairwise syntenic relationships among the full-length *TPS* genes in angiosperms on the phylogenetic tree (Letunic and Bork, 2007). Compared to the TPS-c and TPS-e/f subfamilies, we detected more synteny relationships among genes of angiosperm-specific TPS (a/b/g) subfamilies (Figure 4A). Although

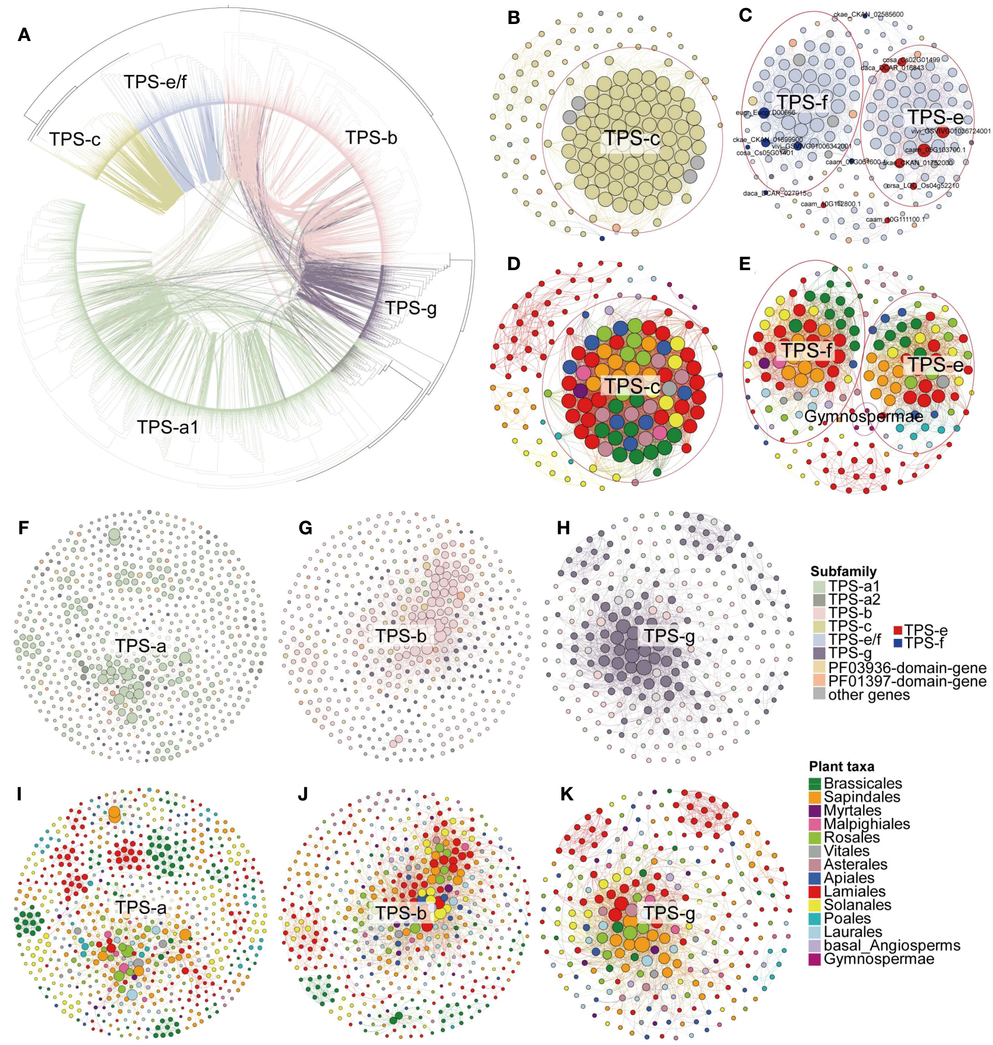

FIGURE 4
Syntenic relationships of full-length *TPS* genes. **(A)** Maximum-likelihood tree of full-length *TPS* genes from angiosperms, depicting their syntenic relationships. Each connecting line inside the circular gene tree represents a syntenic relationship between two genes, with lines color-coded for specific *TPS* subfamilies. **(B, C, F–H)** Synteny subnetworks of TPS-c, TPS-e/f, TPS-a, TPS-b, and TPS-g subfamily, respectively, with node colors (color panel) representing the different gene subfamily assignments. The red and blue nodes in **(C)** indicate TPS-e and TPS-f subfamily genes with established gene classifications (Supplementary Table 4). **(D, E, I–K)** Synteny subnetworks of the TPS-c, TPS-e/f, TPS-a, TPS-b, and TPS-g subfamily, respectively, with node colors denoting the different plant lineages (color panel for species order). Genes circled in red for panels **(B–E)** mark the inter-lineage conserved TPS-c, TPS-e, and TPS-f synteny clusters.

TPS-c and TPS-e/f were closely related in phylogeny, links between genes of these two subfamilies were rare (Figure 4A). Interestingly, in some cases, connections were found between genes from the distal gene clades TPS-e/f and TPS-a/b/g (Figure 4A). One possible interpretation is that TPS-c and TPS-e/f diverged in early land plants, while the most common ancestral gene of TPS-a, TPS-b, and TPS-g expanded after the split of the gymnosperm and angiosperm lineages, resulting in more gene synteny between them. Moreover, 93 PF03936-domain-genes and 42 PF01397-domain-genes shared the synteny relationship with full-length genes (Figure 3B; Supplementary Figure 5; Supplementary Table 14), which implies that these genes might have originated from the domain loss of the full-length genes.

The evolutionary footprint of terpene-related BGCs

We performed a comprehensive inter-specific BGC homology analysis among 74 plants (Supplementary Table 1) to investigate the evolutionary footprint of terpene-related BGCs. We identified a total of 512 terpene-related BGCs, including 380 terpene BGCs, 8 sesterterpene BGCs, and 124 hybrid BGCs (Supplementary Tables 15–17). Among these BGCs, 333 (65.0%) contained full-length *TPS* genes, 97 (9.2%) contained PF03936/PF01397-domain-genes, and 132 (25.8%) were without *TPS* related genes (Supplementary Figure 6; Supplementary Table 18). Through comparative genomic analysis, we have discovered that these *TPS* containing BGCs are either highly conserved across multiple lineages (highly conserved BGCs), conserved within a single lineage (lineage-specific BGCs), or specific to individual species (species-specific BGCs) (Supplementary Tables 19–21). These BGCs distinctly display the footprints of their formation and diversification during plant evolution.

In this study, we chose three highly conserved BGCs and investigated their evolutionary differentiation (Supplementary Table 19). The “vivi_Cluster_7|terpene” BGC was found in *Vitis vinifera* and shares a conserved syntenic block with 17 angiosperm species. Moreover, seven homologous BGCs were found only in the syntenic blocks of Sapindales species (Figure 5A; Supplementary Table 19). The numbers and sequences of *TPS* and *prenyltransferase (PT)* genes in these homologous BGCs show remarkable divergence (Figure 5A; Supplementary Figure 7; Supplementary Table 19). Both *TPS* and *PT* are involved in the terpene synthesis pathway. The “hbra_Cluster_48|terpene” BGC in *Hevea brasiliensis* shares conserved syntenic blocks with 30 angiosperm species, and 14 conserved homologous BGCs were identified in the syntenic blocks of six Sapindales, four Rosales, one Lamiales, and three Solanales species (Figure 6; Supplementary Table 19). All homologous BGCs from Lamiales and Solanales species lost the TPS-g and Epimerase genes (Figure 6; Supplementary Table 19). The “eugr_Cluster_22|terpene” BGC in *Eucalyptus grandis* has a conserved syntenic block with 55 angiosperms and one gymnosperm, and 10 homologous BGCs were identified in four Sapindales, one Malpighiales, and five Rosales species (Supplementary Table 19). In these homologous BGCs, significant TPS-a1 gene tandem duplication and loss of metabolic genes were also observed (Supplementary Table 19).

The unique full-length gene *vivi_GSVIVG01000401001* (*VvTPS34*) within the conserved “vivi_Cluster_7|terpene” terpene BGC in *V. vinifera* was functionally characterized as an (*E*)-β-Ocimene synthase (Martin et al., 2010). (*E*)-β-Ocimene is a commonly observed monoterpene (Fäldt et al., 2003), suggesting this conserved BGC and its homologous BGCs may be involved in the biosynthesis of this monoterpene. Further potential functional prediction of the full-length *TPS* genes in the “eugr_Cluster_22|terpene” BGC aligned to (+)-delta-cadinene synthase in UniprotKB (https://www.uniprot.org/) with 72-100% identity (Supplementary Table 20), indicating this BGC and homologous BGCs may be involved in the biosynthesis of this sesquiterpene. The genes within the conserved “hbra_Cluster_48|terpene” BGC showed no high sequence similarity to any characterized proteins in UniprotKB, precluding functional prediction. Thus, the potential function of this conserved BGC remains unknown.

We observed variations in the number of metabolic genes associated with the divergence of homologous BGCs in the single lineage (Supplementary Table 21). For instance, both the “bnap_Cluster_51|terpene” and “bnap_Cluster_4|terpene” BGCs of *Brassica napus* are sharing synteny blocks with three homologous BGCs of *B. rapa*, *B. juncea* and *A. thaliana* (Figure 5B). Among these homologous BGCs, the number of cytochrome 450 genes is inconsistent, while the polyprenyl synthetase, epimerase, and dioxygenase genes are lost in some BGCs (Figure 5B). Insertion of new metabolic genes and frequent tandem duplications of *TPS* genes, especially the lineage-specific expansion of angiosperms-specific TPS-a/b/g subfamilies, may be associated with the formation of these species-specific BGCs. Examples of these BGCs include the tandem duplication of TPS-d3 and/or Cytochrome 450 genes in “tawa_Cluster_14|terpen” and “tawa_Cluster_48|terpen BGCs”, the TPS-f genes expansion in *Secale cereale*, and TPS-a1 genes expansion in *Ocimum basilicum* (Supplementary Figure 8; Supplementary Table 22).

Discussion

HGT and gene fusion origin of *TPS* genes

HGT of the microbial genes to plants is believed to occur frequently, and has facilitated the transition from aquatic to terrestrial environments for land plants (Ma et al., 2022; Zhang et al., 2022). Gene fusion is an important driver in the evolution of multidomain proteins (Man et al., 2020). It has been reported that HGT and gene fusion may collectively contribute to the origination of full-length *TPS* gene in land plants. Based on extensive microbial sampling and sequence survey, we found that many PF01397 and PF03936 domain sequences shared similarity across different kingdoms, confirming an early origin of these two domains and the potential for multiple instances of gene transfer among microbes and plants.

However, full-length *TPS* gene were very rare or absent in microbes and lower plants. Thus, we concluded that the fusion of

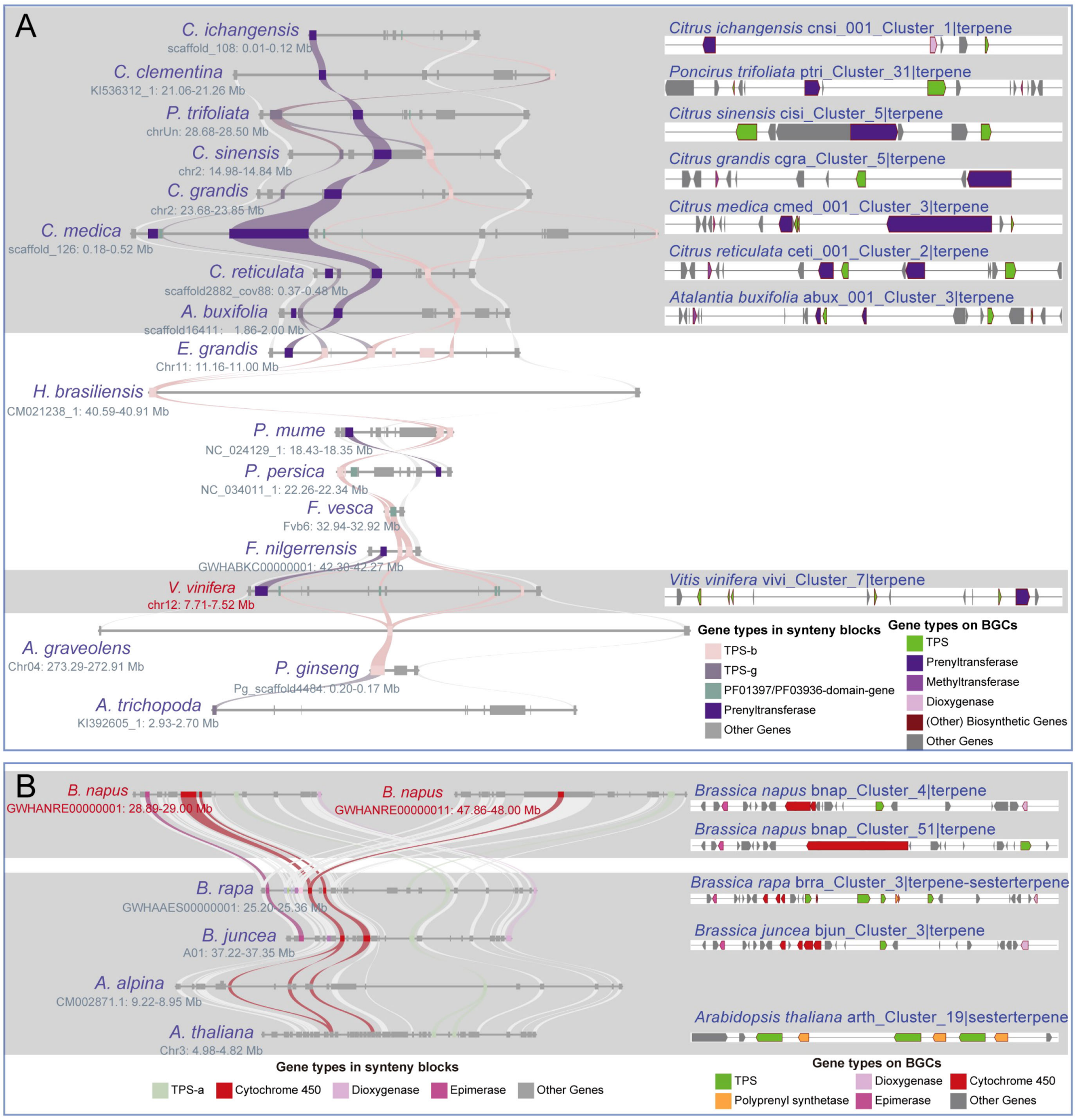
A
C. ichangensis
scaffold_108: 0.01-0.12 Mb
C. clementina
KI536312_1: 21.06-21.26 Mb
P. trifoliata
chrUn: 28.68-28.50 Mb
C. sinensis
chr2: 14.98-14.84 Mb
C. grandis
chr2: 23.68-23.85 Mb
C. medica
scaffold_126: 0.18-0.52 Mb
C. reticulata
scaffold2882_cov88: 0.37-0.48 Mb
A. buxifolia
scaffold16411: 1.86-2.00 Mb
E. grandis
Chr11: 11.16-11.00 Mb
H. brasiliensis
CM021238_1: 40.59-40.91 Mb
P. mume
NC_024129_1: 18.43-18.35 Mb
P. persica
NC_034011_1: 22.26-22.34 Mb
F. vesca
Fvb6: 32.94-32.92 Mb
F. nilgerrensis
GWHABKC00000001: 42.30-42.27 Mb
V. vinifera
chr12: 7.71-7.52 Mb
A. graveolens
Chr04: 273.29-272.91 Mb
P. ginseng
Pg_scaffold4484: 0.20-0.17 Mb
A. trichopoda
KI392605_1: 2.93-2.70 Mb
Citrus ichangensis cnsi_001_Cluster_1|terpene
Poncirus trifoliata ptri_Cluster_31|terpene
Citrus sinensis cisi_Cluster_5|terpene
Citrus grandis cgra_Cluster_5|terpene
Citrus medica cmed_001_Cluster_3|terpene
Citrus reticulata ceti_001_Cluster_2|terpene
Atalantia buxifolia abux_001_Cluster_3|terpene
Vitis vinifera vivi_Cluster_7|terpene
Gene types in synteny blocks
TPS-b
TPS-g
PF01397/PF03936-domain-gene
Prenyltransferase
Other Genes
Gene types on BGCs
TPS
Prenyltransferase
Methyltransferase
Dioxygenase
(Other) Biosynthetic Genes
Other Genes
B
B. napus
GWHANRE00000001: 28.89-29.00 Mb
B. napus
GWHANRE00000011: 47.86-48.00 Mb
B. rapa
GWHAAES00000001: 25.20-25.36 Mb
B. juncea
A01: 37.22-37.35 Mb
A. alpina
CM002871.1: 9.22-8.95 Mb
A. thaliana
Chr3: 4.98-4.82 Mb
Brassica napus bnap_Cluster_4|terpene
Brassica napus bnap_Cluster_51|terpene
Brassica rapa brra_Cluster_3|terpene-sesterterpene
Brassica juncea bjun_Cluster_3|terpene
Arabidopsis thaliana arth_Cluster_19|sesterterpene
Gene types in synteny blocks
TPS-a
Cytochrome 450
Dioxygenase
Epimerase
Other Genes
Gene types on BGCs
TPS
Dioxygenase
Cytochrome 450
Polyprenyl synthetase
Epimerase
Other Genes

FIGURE 5
Differentiation of the homologous BGCs. **(A)** The "vivi_Cluster_7|terpene" BGC in *V. vinifera* shares synteny blocks with 17 angiosperms from 6 lineages, and 7 homologous BGCs are identified within seven Sapindales species. **(B)** The "bnap_Cluster_4|terpene" and "bnap_Cluster_4|terpene" BGCs share synteny blocks with 4 Brassicales species, and 3 homologous BGCs are identified within these synteny blocks. Both **(A, B)** demonstrate homologous BGCs within varying differentiation.

PF03936 (C-terminal) and PF01397 (N-terminal) domains more likely occurred in the ancestral land plant after the divergence from lower plants, and these two domains could have been individually acquired from microbe via HGT. Our sequence analysis found one case that a potential full-length *TPS* gene may derive by HGT from plants to soil bacteria, suggesting the HGT of *TPS* is possibly also a pathway for terpene innovation in microbes.

Conservation and differentiation dynamics of *TPS* subfamilies

Terpenes involved in primary metabolism are synthesized through relatively conserved pathways in plants (Yang et al., 2020). Our results reveal that the TPS-c and TPS-e/f subfamilies, which are involve in GA production, display high gene synteny conservation among angiosperms with small-scale lineage-specific expansions. In contrast, the angiosperm-specific TPS-a/b/g subfamilies, dedicated to secondary metabolism, exhibit significant lineage-specific patterns. *TPS* genes possess profound functional plasticity, where minor changes in active sites can dramatically affect catalytic properties, allowing for the emergence of new functions with minimal investment in evolving new enzymes (Karunanithi and Zerbe, 2019). The evolution of the ability to synthesize specialized metabolites has likely been crucial for the survival and diversification of various plant species (Qi et al., 2004). Consequently, the expansion and subfunctionalization/neofunctionalization of lineage-specific *TPS* genes facilitate the

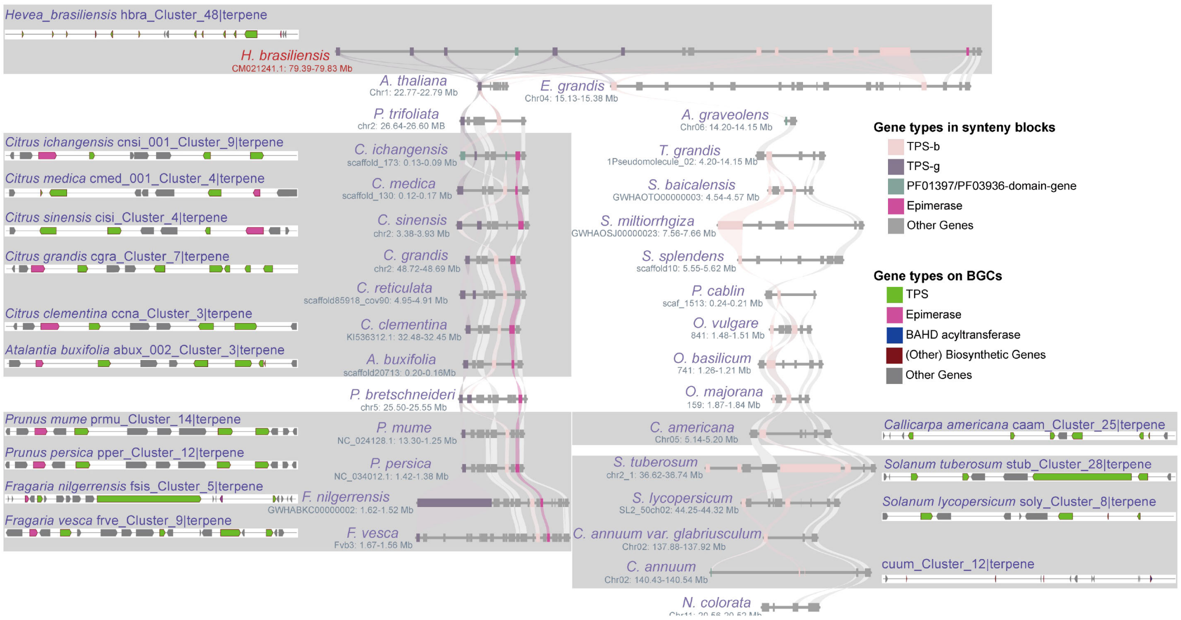
Hevea_brasiliensis hbra_Cluster_48|terpene
H. brasiliensis
CM021241.1: 79.39-79.83 Mb
A. thaliana
Chr1: 22.77-22.79 Mb
E. grandis
Chr04: 15.13-15.38 Mb
P. trifoliata
chr2: 26.64-26.60 MB
Citrus ichangensis cnsi_001_Cluster_9|terpene
C. ichangensis
scaffold_173: 0.13-0.09 Mb
Citrus medica cmed_001_Cluster_4|terpene
C. medica
scaffold_130: 0.12-0.17 Mb
Citrus sinensis cisi_Cluster_4|terpene
C. sinensis
chr2: 3.38-3.93 Mb
Citrus grandis cgra_Cluster_7|terpene
C. grandis
chr2: 48.72-48.69 Mb
C. reticulata
scaffold85918_cov90: 4.95-4.91 Mb
Citrus clementina ccna_Cluster_3|terpene
C. clementina
KI536312.1: 32.48-32.45 Mb
Atalantia buxifolia abux_002_Cluster_3|terpene
A. buxifolia
scaffold20713: 0.20-0.16Mb
P. bretschneideri
chr5: 25.50-25.55 Mb
Prunus mume prmu_Cluster_14|terpene
P. mume
NC_024128.1: 13.30-1.25 Mb
Prunus persica pper_Cluster_12|terpene
P. persica
NC_034012.1: 1.42-1.38 Mb
Fragaria nilgerrensis fsis_Cluster_5|terpene
F. nilgerrensis
GWHABKC00000002: 1.62-1.52 Mb
Fragaria vesca frve_Cluster_9|terpene
F. vesca
Fvb3: 1.67-1.56 Mb
A. graveolens
Chr06: 14.20-14.15 Mb
T. grandis
1Pseudomolecule_02: 4.20-14.15 Mb
S. baicalensis
GWHAOTO00000003: 4.54-4.57 Mb
S. miltiorrhgiza
GWHAOSJ00000023: 7.56-7.66 Mb
S. splendens
scaffold10: 5.55-5.62 Mb
P. cablin
scaf_1513: 0.24-0.21 Mb
O. vulgare
841: 1.48-1.51 Mb
O. basilicum
741: 1.26-1.21 Mb
O. majorana
159: 1.87-1.84 Mb
C. americana
Chr05: 5.14-5.20 Mb
Callicarpa americana caam_Cluster_25|terpene
S. tuberosum
chr2_1: 36.62-36.74 Mb
Solanum tuberosum stub_Cluster_28|terpene
S. lycopersicum
SL2_50ch02: 44.25-44.32 Mb
Solanum lycopersicum soly_Cluster_8|terpene
C. annuum var. glabriusculum
Chr02: 137.88-137.92 Mb
C. annuum
Chr02: 140.43-140.54 Mb
cuum_Cluster_12|terpene
N. colorata
Gene types in synteny blocks
TPS-b
TPS-g
PF01397/PF03936-domain-gene
Epimerase
Other Genes
Gene types on BGCs
TPS
Epimerase
BAHD acyltransferase
(Other) Biosynthetic Genes
Other Genes

FIGURE 6
Differentiation of homologous BGCs. The “hbra_Cluster_48|terpene” BGC in *H. brasiliensis* shares synteny blocks with 31 angiosperms from 19 lineages, and its 15 homologous BGCs were identified within the synteny blocks of Sapindales, Rosales, Lamiales, and Solanales species, demonstrating the differentiation of homologous BGCs in terms of both the number and type of metabolic genes.

production of various terpene metabolites in response to changing biotic and abiotic environments, contributing to the differentiation of plant ecotypes (Tholl, 2015; Pichersky and Raguso, 2018; Jiang et al., 2019). The diversity of PVTs also leads to significantly variable compositions of oils produced in various plants.

Angiosperms have undoubtedly dominated recent ecological history on land by frequent gene duplication (Zhang et al., 2022), polyploidy events, chromosome reorganization, and other molecular mechanisms that create a rich source of genetic novelty and adaptation to complex environments. These activities around genome structural variations may have contributed to the significant expansion and lineage-specific diversification of TPS-a/b/g subfamilies in angiosperms. Phylogenetic studies also show the deep divergence between angiosperm-specific subfamily TPS-a/b/g and other four subfamilies (TPS-c, TPS-e/f, TPS-d, TPS-h). Thus, the subfunctionalization/neofunctionalization of TPS-a/b/g subfamilies in various angiosperms lineages led to a diversity of enzymes, contributing to significant terpene and species phenotypic diversity.

Gain and loss of protein domains are widespread evolutionary events (Man et al., 2020) that can generate novel genetic diversity and further species diversity. By revealing the relationships of the PF03936/PF01397-domain-genes and full-length genes in plants, this study sheds light on the complex evolutionary trajectories of plant *TPSs* and how this domain gain and loss has been involved in shaping the terpene chemical diversity. However, it should be noted these findings are susceptible to gene annotation errors, which must be carefully considered in future studies. The broad distribution of PF03936/PF01397-domain-genes across all *TPS* subfamily clades, along with their syntenic relationships with full-length *TPS* genes, supports the hypothesis that these two domains can be lost in some full-length *TPS* genes. In addition to domain shuffling and loss, the outright loss of entire subfamilies during genome evolution of individual species represents another mechanism contributing to the inter-specific diversity of terpene profiles.

Terpene-related BGCs

The establishment and maintenance of BGCs has been driven by natural selection, including long-term purifying selection, positive selection, and balancing selection (Wu et al., 2022). Thus, the conserved BGCs imply a stronger selective advantage in the genome (Kautsar et al., 2017). In this study, we identified three conserved terpene BGCs as highly conserved synteny blocks that are shared among multiple plant lineages, suggesting that secondary metabolic terpenes produced by these BGCs may play a crucial role in defense or interactions with the external environment of these plant lineages. There is compelling evidence that BGCs in plants arose from gene duplication, neofunctionalization, genomic relocation, and chromosomal inversion (Nützmann et al., 2018; Rokas et al., 2018; Liu et al., 2020; Wu et al., 2022). We found that gene tandem duplication, loss, or gain of metabolic genes were associated with the divergence of homologous BGCs. Moreover, lineage-specific expansion of *TPS* genes and the insertion of new metabolic genes are probably the primary pathways for the formation of species-specific BGCs. These homologous or specific

BGCs enrich the diversity of plant terpene metabolites and facilitate plants' survival advantage in different ecological niches. However, further investigation into the functions of these BGCs are needed to better understand the relationship between BGCs and plant terpene metabolites.

This study employed comprehensive phylogenetic samplings and analysis of terpene synthase sequences from plants and microbes to elucidate the origin and diversification of *TPS* genes in land plants. We revealed gene fusion of PF01397 and PF03936 domain occurred in early land plants, and the synteny relationships and diversification of *TPS* genes across major land plant groups. Additionally, we identified plant *TPS* gene synteny relationships and the evolution of homologous and species-specific terpene-relate BGCs. Our study thus provides new and insightful perspectives into the diversity of terpene biosynthesis in plants and presents invaluable resources for future evolutionary and functional studies. This could also enhance our understanding and further inform the biotechnology industry on how to optimize the utilization of oil-producing plants.

Statistics and reproducibility

No statistical tests were employed in this study. All methods used for sequence analyses are described in the corresponding methods.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: All single domain and full-length *TPS* sequences obtained from various microbes and plants and the sequence alignments (PF01397 domain, PF03936 domain and full-length gene sequences) could be accessed on GitHub at the following link: https://github.com/xmy-1682/TPS-data/tree/master/sequences_data and https://github.com/xmy-1682/TPS-data/tree/master/Sequence_alignments. The Newick In review format of all phylogenetic trees (in Figures 2, 3, Supplementary Figures 1, 2) generated in this study are available on GitHub at: https://github.com/xmy-1682/TPSdata/tree/master/Newick_format_of_phylogenetic_trees. The raw data used for regenerating the gene synteny network can be found at: https://github.com/xmy 1682/TPS-data/tree/master/Raw_data. All terpene-related BGCs identified in 72 plant species are listed in Supplementary Table 15, and could be accessed on GitHub at: https://github.com/xmy-1682/TPS-data/tree/master/PlantiSMASH_results].

Author contributions

J-FM: Methodology, Supervision, Writing – review & editing, Writing – original draft. X-MY: Data curation, Formal Analysis, Investigation, Methodology, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing. S-SZ: Investigation, Visualization, Writing – review & editing. HL: Investigation, Visualization, Writing – review & editing. S-WZ: Visualization, Writing – review & editing. X-CT: Visualization, Writing – review & editing. T-LS: Visualization, Writing – review & editing. Y-TB: Visualization, Writing – review & editing. Z-CL: Visualization, Writing – review & editing. K-HJ: Visualization, Writing – review & editing. SN: Visualization, Writing – review & editing. J-FG: Visualization, Writing – review & editing. LK: Visualization, Writing – review & editing. IP: Investigation, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was supported by the National Key R&D Program of China (2022YFD2200103) and National Natural Science Foundation of China (32171816).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1273648/full#supplementary-material

References

Alquézar, B., Rodríguez, A., de la Peña, M., and Peña, L. (2017). Genomic analysis of terpene synthase family and functional characterization of seven sesquiterpene synthases from. *Citrus sinensis Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01481

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208. doi: 10.1093/nar/gkp335

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. Proceedings of the international AAAI conference on web and social media 3 (1), 361–362. doi: 10.1609/icwsm.v3i1.13937

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., et al. (2004). The Pfam protein families database. *Nucleic Acids Res.* 32, D138–D141. doi: 10.1093/nar/gkh121

Bohlmann, J., Meyer-Gauen, G., and Croteau, R. (1998). Plant terpenoid synthases: Molecular biology and phylogenetic analysis. *Proc. Natl. Acad. Sci.* 95, 4126. doi: 10.1073/pnas.95.8.4126

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

Cao, R., Zhang, Y., Mann, F. M., Huang, C., Mukkamala, D., Hudock, M. P., et al. (2010). Diterpene cyclases and the nature of the isoprene fold. *Proteins* 78, 2417–2432. doi: 10.1002/prot.22751

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. doi: 10.1093/bioinformatics/btp348

Carson, C. F., Hammer, K. A., and Riley, T. V. (2006). Melaleuca alternifolia (Tea Tree) oil: a review of antimicrobial and other medicinal properties. *Clin. Microbiol. Rev.* 19, 50–62. doi: 10.1128/cmr.19.1.50-62.2006

Chaw, S.-M., Liu, Y. -C., Wu, Y. -W., Wang, Y. -H., Lin, C. -Y. I., Wu, C. -S., et al. (2019). Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution. *Nat. Plants* 5, 63–73. doi: 10.1038/s41477-018-0337-0

Chen, G., Xie, Y., Zhou, D., Yang, Y., Liu, J., Hou, Y., et al. (2020). Chemical constituents from shells of Xanthoceras sorbifolium. *Phytochemistry* 172, 112288. doi: 10.1016/j.phytochem.2020.112288

Chen, Z., Vining, K. J., Qi, X., Yu, X., Zheng, Y., Liu, Z., et al. (2021). Genome-wide analysis of terpene synthase gene family in *Mentha longifolia* and catalytic activity analysis of a single terpene synthase. *Genes* 12, 518. doi: 10.3390/genes12040518

Chen, F., Tholl, D., Bohlmann, J., and Pichersky, E. (2011). The family of terpene synthases in plants: A mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant J.* 66, 212–229. doi: 10.1111/j.1365-313X.2011.04520.x

Cheng, S., Xian, W., Fu, Y., Marin, B., Keller, J., Wu, T., et al. (2019). Genomes of subaerial zygnematophyceae provide insights into land plant evolution. *Cell* 179, 1057–1067.e1014. doi: 10.1016/j.cell.2019.10.019

Derényi, I., Palla, G., and Vicsek, T. (2005). Clique percolation in random networks. *Phys. Rev. Lett.* 94, 160202. doi: 10.1103/PhysRevLett.94.160202

Fäldt, J., Arimura, G.-i., Gershenzon, J., Takabayashi, J., and Bohlmann, J. (2003). Functional identification of *AtTPS03* as (E)-*β*-ocimene synthase: a monoterpene synthase catalyzing jasmonate- and wound-induced volatile formation in Arabidopsis thaliana. *Planta* 216, 745–751. doi: 10.1007/s00425-002-0924-0

Fortunato, S. (2010). Community detection in graphs. *Phys. Rep.* 486, 75–174. doi: 10.1016/j.physrep.2009.11.002

Hayashi, K.-i., Kawaide, H., Notomi, M., Sakigi, Y., Matsuo, A., Nozaki, H., et al. (2006). Identification and functional analysis of bifunctional *ent*-kaurene synthase from the moss *Physcomitrella patens*. *FEBS Lett.* 580, 6175–6181. doi: 10.1016/j.febslet.2006.10.018

Jia, Q., Brown, R., Köllner, T. G, Fu, J., Chen, X., Wong, G. K. S., et al. (2022). Origin and early evolution of the plant terpene synthase family. *Proc. Natl. Acad. Sci.* 119, e2100361119. doi: 10.1073/pnas.2100361119

Jia, Q., Köllner, T. G., Gershenzon, J., and Chen, F. (2018). MTPSLs: New terpene synthases in nonseed plants. *Trends Plant Sci.* 23, 121–128. doi: 10.1016/j.tplants.2017.09.014

Jiang, S.-Y., Jin, J., Sarojam, R., and Ramachandran, S. (2019). A comprehensive survey on the terpene synthase gene family provides new insight into its evolutionary patterns. *Genome Biol. Evol.* 11, 2078–2098. doi: 10.1093/gbe/evz142

Kainat, R., Mushtaq, Z., and Nadeem, F. (2019). Derivatization of essential oil of Eucalyptus to obtain valuable market products-A comprehensive review. *International Journal of Chemical and Biochemical Sciences* 15, 58–68.

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. doi: 10.1038/nmeth.4285

Karunanithi, P. S., and Zerbe, P. (2019). Terpene synthases as metabolic gatekeepers in the evolution of plant terpenoid chemical diversity. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.01166

Katoh, K., Misawa, K., Kuma, K. i., and Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/nar/gkf436

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. doi: 10.1093/molbev/mst010

Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A., and Medema, M. H. (2017). plantiSMASH: Automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.* 45, W55–W63. doi: 10.1093/nar/gkx305

Kawaide, H., Hayashi, K. -i., Kawanabe, R., Sakigi, Y., Matsuo, A., Natsume, M., et al. (2011). Identification of the single amino acid involved in quenching the *ent*-kauranyl cation by a water molecule in *ent*-kaurene synthase of *Physcomitrella patens*. *FEBS J.* 278, 123–133. doi: 10.1111/j.1742-4658.2010.07938.x

King, D. J., Gleadow, R. M., and Woodrow, I. E. (2006). Regulation of oil accumulation in single glands of Eucalyptus polybractea. *New Phytol.* 172, 440–451. doi: 10.1111/j.1469-8137.2006.01842.x

Köksal, M., Jin, Y., Coates, R. M., Croteau, R., and Christianson, D. W. (2011). Taxadiene synthase structure and evolution of modular architecture in terpene biosynthesis. *Nature* 469, 116–120. doi: 10.1038/nature09628

Letunic, I., and Bork, P. (2007). Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics* 23, 127–128. doi: 10.1093/bioinformatics/btl529

Li, R., Wang, K., Wang, D., Xu, L., Shi, Y., Dai, Z., et al. (2021). Production of plant volatile terpenoids (rose oil) by yeast cell factories. *Green Chem.* 23, 5088–5096. doi: 10.1039/D1GC00917F

Little, D. B., and Croteau, R. B. (1999). Biochemistry of essential oil terpene. In Teranishi, R., Wick, E. L., and Hornstein, I. (eds). *Flavor Chemistry*. (Boston, MA: Springer), 239–253. doi: 10.1007/978-1-4615-4693-1_21

Liu, Z., Cheema, J., Vigouroux, M., Hill, L., Reed, J., Paajanen, P., et al. (2020). Formation and diversification of a paradigm biosynthetic gene cluster in plants. *Nat. Commun.* 11, 5354. doi: 10.1038/s41467-020-19153-6

Ma, J., Wang, S., Zhu, X., Sun, G., Chang, G., Li, L., et al. (2022). Major episodes of horizontal gene transfer drove the evolution of land plants. *Mol. Plant* 15, 857–871. doi: 10.1016/j.molp.2022.02.001

Man, J., Gallagher, J. P., and Bartlett, M. (2020). Structural evolution drives diversification of the large LRR-RLK gene family. *New Phytol.* 226, 1492–1505. doi: 10.1111/nph.16455

Martin, D. M., Aubourg, S., Schouwey, M. B., Daviet, L., Schalk, M., Toub, O., et al. (2010). Functional annotation, genome organization and phylogeny of the grapevine (*Vitis vinifera*) terpene synthase gene family based on genome assembly, FLcDNA cloning, and enzyme assays. *BMC Plant Biol.* 10, 226. doi: 10.1186/1471-2229-10-226

McGarvey, D. J., and Croteau, R. (1995). Terpenoid metabolism. *Plant Cell* 7, 1015–1026. doi: 10.1105/tpc.7.7.1015

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* 41, e121–e121. doi: 10.1093/nar/gkt263

Morrone, D., Chambers, J., Lowry, L., Kim, G., Anterola, A., Bender, K., et al. (2009). Gibberellin biosynthesis in bacteria: Separate *ent*-copalyl diphosphate and *ent*-kaurene synthases in *Bradyrhizobium japonicum*. *FEBS Lett.* 583, 475–480. doi: 10.1016/j.febslet.2008.12.052

Naidoo, S., Külheim, C., Zwart, L., Mangwanda, R., Oates, C. N., Visser, E. A., et al. (2014). Uncovering the defence responses of Eucalyptus to pests and pathogens in the genomics age. *Tree Physiol.* 34, 931–943. doi: 10.1093/treephys/tpu075

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2014). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Nützmann, H.-W., Scazzocchio, C., and Osbourn, A. J. A. R. (2018). o. G. Metabolic gene clusters in eukaryotes. *Annu. Rev. Genet.* 52, 159–183. doi: 10.1146/annurev-genet-120417-031237

Palla, G., Derényi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435, 814–818. doi: 10.1038/nature03607

Pichersky, E., and Raguso, R. A. (2018). Why do plants produce so many terpenoid compounds? *New Phytol.* 220, 692–702. doi: 10.1111/nph.14178

Qi, X., Bakht, S., Leggett, M., Maxwell, C., Melton, R., Osbourn, A., et al. (2004). A gene cluster for secondary metabolism in oat: Implications for the evolution of metabolic diversity in plants. *Proc. Natl. Acad. Sci.* 101, 8233–8238. doi: 10.1073/pnas.0401301101

Rokas, A., Wisecaver, J. H., and Lind, A. L. (2018). The birth, evolution and death of metabolic gene clusters in fungi. *Nat. Rev. Microbiol.* 16, 731–744. doi: 10.1038/s41579-018-0075-3

Rosvall, M., and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci.* 105, 1118–1123. doi: 10.1073/pnas.0706851105

Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PloS One* 11(10):e0163962. doi: 10.1371/journal.pone.0163962

Shen, W., and Ren, H. (2021). TaxonKit: A practical and efficient NCBI taxonomy toolkit. *J. Genet. Genomics* 48, 844–850. doi: 10.1016/j.jgg.2021.03.006

Smanski, M. J., Peterson, R. M., Huang, S.-X., and Shen, B. (2012). Bacterial diterpene synthases: new opportunities for mechanistic enzymology and engineered biosynthesis. *Curr. Opin. Chem. Biol.* 16, 132–141. doi: 10.1016/j.cbpa.2012.03.002

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

The Angiosperm Phylogeny, G., Chase, M. W., Christenhusz, M. J. M., Fay, M. F., Byng, J. W., Judd, W. S., et al. (2016). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot J. Linn. Soc.* 181, 1–20. doi: 10.1111/boj.12385

Tholl, D. (2015). Biosynthesis and biological functions of terpenoids in plants. *Biotechnol. Isoprenoids* 148:63–106. doi: 10.1007/10_2014_295

Trapp, S., and Croteau, R. (2001). Defensive resin biosynthesis in conifers. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 52, 689–724. doi: 10.1146/annurev.arplant.52.1.689

Vining, K. J., Johnson, S. R., Ahkami, A., Lange, I., Parrish, A. N., Trapp, S. C., et al. (2017). Draft genome sequence of mentha longifolia and development of resources for mint cultivar improvement. *Mol. Plant* 10, 323–339. doi: 10.1016/j.molp.2016.10.018

Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40, e49. doi: 10.1093/nar/gkr1293

Wang, D., Su, D., Yu, B., Chen, C., Cheng, L., Li, X., et al. (2017). Novel anti-tumour barringenol-like triterpenoids from the husks of Xanthoceras sorbifolia Bunge and their three dimensional quantitative structure activity relationships analysis. *Fitoterapia* 116, 51–60. doi: 10.1016/j.fitote.2016.11.002

Witjes, L., Kooke, R., van der Hooft, J. J. J., de Vosa, R. C. H., Keurentjes, J. J. B., Medema, M. H., et al. (2019). A genetical metabolomics approach for bioprospecting plant biosynthetic gene clusters. *BMC Res. Notes* 12, 194. doi: 10.1186/s13104-019-4222-3

Wu, D., Jiang, B., Ye, C.-Y., Timko, M. P., and Fan, L. (2022). Horizontal transfer and evolution of the biosynthetic gene cluster for benzoxazinoid in plants. *Plant Commun.* 3, 100320. doi: 10.1016/j.xplc.2022.100320

Yang, M., Liu, G., Yamamura, Y., Chen, F., and Fu, J. (2020). Divergent evolution of the diterpene biosynthesis pathway in tea plants (*Camellia sinensis*) caused by single amino acid variation of *ent*-kaurene synthase. *J. Agric. Food Chem.* 68, 9930–9939. doi: 10.1021/acs.jafc.0c03488

Yu, Z., Zhao, C., Zhang, G., Teixeira da Silva, J. A., and Duan, J. (2020). Genome-wide identification and expression profile of TPS gene family in *Dendrobium officinale* and the role of *DoTPS10* in linalool biosynthesis. *Int. J. Mol. Sci.* 21, 5419. doi: 10.3390/ijms21155419

Zhang, L., Wu, S., Chang, X., Wang, X., Zhao, Y., Xia, Y., et al. (2020). The ancient wave of polyploidization events in flowering plants and their facilitated adaptation to environmental stress. *Plant Cell Environ.* 43, 2847–2856. doi: 10.1111/pce.13898

Zhang, Z., Ma, X., Liu, Y., Yang, L., Shi, X., Wang, H., et al. (2022). Origin and evolution of green plants in the light of key evolutionary events. *J. Integr. Plant Biol.* 64, 516–535. doi: 10.1111/jipb.13224

Zhao, T., Holmer, R., de Bruijn, S., Angenent, G. C., van den Burg, H. A., Schranz, M. E., et al. (2017). Phylogenomic synteny network analysis of MADS-box transcription factor genes reveals lineage-specific transpositions, ancient tandem duplications, and deep positional conservation. *Plant Cell* 29, 1278. doi: 10.1105/tpc.17.00312

Zhao, T., and Schranz, M. E. (2019). Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. *Proc. Natl. Acad. Sci.* 116, 2165. doi: 10.1073/pnas.1801757116

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

Zhou, F., and Pichersky, E. (2020). The complete functional characterisation of the terpene synthase family in tomato. *New Phytol.* 226, 1341–1360. doi: 10.1111/nph.16431

Zi, J., Mafu, S., and Peters, R. J. (2014). To gibberellins and beyond! Surveying the evolution of (di)terpenoid metabolism. *Annu. Rev. Plant Biol.* 65, 259–286. doi: 10.1146/annurev-arplant-050213-035705