the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2022) **112**, 630–645 doi: 10.1111/tpj.15968

Chromosome-level genome assembly and functional characterization of terpene synthases provide insights into the volatile terpenoid biosynthesis of *Wurfbainia villosa*

Peng Yang[1,2,3], Hai-Ying Zhao[2,4], Jie-Shu Wei[5], Yuan-Yuan Zhao[2], Xiao-Jing Lin[1,2], Jing Su[6], Fang-Ping Li[7], Meng Li[2], Dong-Ming Ma[1,2], Xu-Kai Tan[8], Hui-Lin Liang[1,2], Ye-Wen Sun[1,2], Ruo-Ting Zhan[1,2], Guo-Zhen He[1,2,*], Xiao-Fan Zhou[7,*] and Jin-Fen Yang[1,2,*]

[1]*School of Pharmaceutical Science, Guangzhou University of Chinese Medicine, Guangzhou 510006, China,*
[2]*Key Laboratory of Chinese Medicinal Resource from Lingnan (Ministry of Education), Guangzhou University of Chinese Medicine, Guangzhou 510006, China,*
[3]*Hunan Provincial Key Laboratory for Synthetic Biology of Traditional Chinese Medicine, School of Pharmaceutical Sciences, Hunan University of Medicine, Huaihua 418000, China,*
[4]*The Second Clinical Medical College of Guangxi University of Science and Technology, Louzhou 5450000, China,*
[5]*School of Pharmacy, Guangzhou Xinhua University, Guangzhou 510520, China,*
[6]*Agricultural Experimental Station of Yangchun City (Amomum villosum Testing farm of Yangchun City), Yangchun 529600, China,*
[7]*Guangdong Province Key Laboratory of Microbial Signals and Disease Control, Integrative Microbiology Research Centre, South China Agricultural University, Guangzhou 510642, China, and*
[8]*Grandomics Biosciences, Beijing 102200, China*

Received 15 February 2022; accepted 4 September 2022; published online 7 September 2022.
*For correspondence (e-mail heguozhen@gzucm.edu.cn; xiaofan_zhou@scau.edu.cn; yangjf@gzucm.edu.cn).

SUMMARY

Wurfbainia villosa **is a well-known medicinal and edible plant that is widely cultivated in the Lingnan region of China. Its dried fruits (called *Fructus Amomi*) are broadly used in traditional Chinese medicine for curing gastrointestinal diseases and are rich in volatile terpenoids. Here, we report a high-quality chromosome-level genome assembly of *W. villosa* with a total size of approximately 2.80 Gb, 42 588 protein-coding genes, and a very high percentage of repetitive sequences (87.23%). Genome analysis showed that *W. villosa* likely experienced a recent whole-genome duplication event prior to the *W. villosa–Zingiber officinale* divergence (approximately 11 million years ago), and a recent burst of long terminal repeat insertions afterward. The *W. villosa* genome enabled the identification of 17 genes involved in the terpenoid skeleton biosynthesis pathway and 66 terpene synthase (TPS) genes. We found that tandem duplication events have an important contribution to the expansion of *WvTPSs*, which likely drove the production of volatile terpenoids. In addition, functional characterization of 18 WvTPSs, focusing on the TPS-a and TPS-b subfamilies, showed that most of these WvTPSs are multi-product TPS and are predominantly expressed in seeds. The present study provides insights into the genome evolution and the molecular basis of the volatile terpenoids diversity in *W. villosa*. The genome sequence also represents valuable resources for the functional gene research and molecular breeding of *W. villosa*.**

Keywords: *Wurfbainia villosa*, chromosome-level genome, terpene synthase, volatile terpenoid biosynthesis, nanopore sequencing, Hi-C.

INTRODUCTION

Wurfbainia villosa (2*n* = 48, homotypic synonym: *Amomum villosum*), a perennial herb plant that belongs to the monophyletic genus *Wurfbainia* of the family Zingiberaceae, has been used in medicine for at least 1300 years, mainly for the treatment of gastrointestinal diseases (Hugo et al., 2018; Chen & Chen, 1982). The dried fruit of *W. villosa* is referred to as *Fructus Amomi* (Chinese medicine name: Sharen), which is one of the famous 'Four Major Southern China Medicines' and plays important roles in the clinical treatment of warming the spleen, eliminating dampness and the prevention of miscarriage diseases (Ke & Shi, 2012). In addition, *Fructus*

Amomi has been approved by the China Food and Drug Administration as a medicine food homology species in China, and it has been widely used in the production of food, liquors, and tea, as well as cosmetics and food additives. Modern studies have demonstrated that the volatile terpenoids of *W. villosa* have a wide range of pharmacological effects, such as anticancer, anti-inflammatory, antimicrobial, and hypoglycemic effects (Chen et al., 2018a,b; Yue et al., 2021; Tang et al., 2021). It has been reported that seeds of the *W. villosa* are rich in volatile terpenoids, and their main bioactive substances are bornyl acetate, borneol, and camphor (Chen et al., 2020a,b,c), and it is worth studying the biosynthesis and organ-specific enrichment mechanisms further.

In (National Pharmacopoeia Committee, 2020), *Fructus Amomi* refers to the dried ripe fruits of three ginger plants, including *W. villosa* (*Amomum villosum*), *Amomum villosum* Lour. var. xanthioides, and *Amomum longiligulare*. Among them, the authentic *W. villosa* produced in Yangchun, Guangdong Province is well known for its high content of volatile oils (Ao et al., 2019), the price of which is five to ten times higher than non-authentic ones in the market in China. The gynandrium-like structure of the flowers of *W. villosa* makes pollinations by insects difficult, leading to low natural fruiting rates, and hand pollinations are usually required to increase the yield (Tang et al., 2012; Yang et al., 2021, 2022). This pollination characteristic of *W. villosa* poses a severe limitation on its yield, which is insufficient to meet the market demand and hinders the development of the industry. Thus, there is an urgent need to breed varieties of high yield and high quality. However, the genome sequence of *W. villosa* has not yet been reported, which restricts the development of functional genomics and molecular breeding of this plant.

The genomic information of *W. villosa* can lay the foundation for improving the quality of medicinal materials, discovering functional genes, accelerating molecular breeding, and protecting wild resources. To this end, in the present study, we report a high-quality chromosome-level reference genome of *W. villosa* by combining Oxford Nanopore Technologies (ONT) sequencing and Hi-C technology. Based on the homolog searching and functional annotations, 66 candidate *WvTPSs* (terpene synthase) are identified. In addition, the functional characterization of 18 *WvTPSs* has been performed to reveal the genetic basis for volatile terpenoids enrichment in *W. villosa* seeds. In conclusion, the present study provides insights into the diversity of volatile terpenoids in *W. villosa*, and also provides genomic resources to facilitate the genetic improvement of this medicinal plant and future investigations of the evolution of Zingiberaceae.

RESULTS

Genome assembly and annotation

The genome of *W. villosa* was sequenced using both the Illumina NovaSeq 6000 and the ONT PromethION high-throughput sequencing platforms, resulting in 163.91 Gb (546.36 million pairs of 150 bp reads) of short-read and 322.69 Gb (14.49 million reads; N50: 31.43 kb) of long-read clean sequencing data, respectively (Table S1). The *W. villosa* genome was estimated to have a size of 2644.9 Mb and a relatively low heterozygosity level of 0.4% based on K-mer analysis of the Illumina short-read sequencing data (Figure S1 and Table S2). A *de novo* assembly of the ONT long-read sequencing data (estimated coverage of approximately 122.2×) was performed with NextDenovo (Table S1), giving rise to a draft assembly of 2799.20 Mb consisting of 1110 contigs (contig N50 value: 9.13 Mb) (Table 1 and Table S3). We then generated 306.93 Gb (1023.09 million pairs of 150 bp read) of Illumina short-read Hi-C data to construct chromosome-level genome assembly. As a result, 826 contigs accounting for about 92.01% of all sequences were anchored into 24 pseudochromosomes with sizes ranging from 37.33 Mb to 139.36 Mb (Table S4). Finally, we obtained a chromosome-level genome of *W. villosa* containing 24 chromosomes with a total size of 2.80 Gb (Figure 1b and Table 1).

We evaluated the quality of the genome assembly using multiple methods. First, the Hi-C interaction heatmap clearly showed that the clustering, ordering, and orientation of contigs are reliable (Figure S2). Second, Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis showed that 97.9 and 0.9% of the 1614 Embryophyta-wide conserved genes are present in the *W. villosa* genome as 'complete' and 'fragmented'

Table 1 Major indicators of the *W. villosa* genome

Assembly features	
Total genome size (Mb)	2799.2
Contig N50 (Mb)	9.1
Contig number	1100
Total scaffolds length (Mb)	2575.5
Scaffold N50 (Mb)	109.9
Pseudochromosomes	24
GC content (%)	40.2%
Complete BUSCO (%)	97.9%
LAI score	16.18
Annotation features	
Number of protein-coding genes	42 588
Average gene length (bp)	5277
Average CDS length (bp)	1192
Percentage of repeat sequences (%)	87.23
Number of ncRNA	7087
Number of rRNA	522
Number of tRNA	1843
Complete BUSCO (%)	95.2%

(a)

(b)

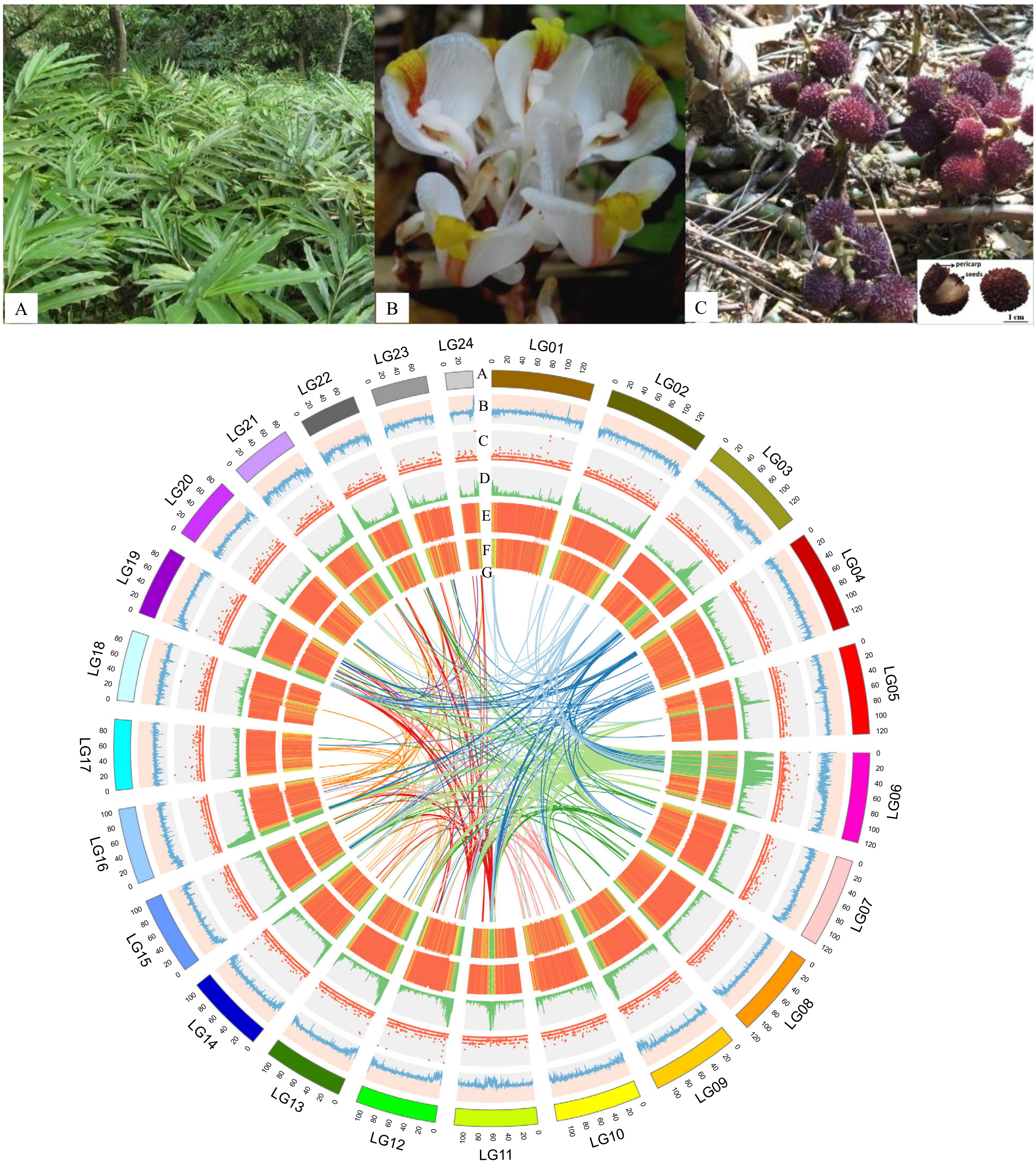

Figure 1. Morphology of *W. villosa* and overview of the *W. villosa* genome assembly. (a) Morphological characteristics of *W. villosa*. A, plant; B, inflorescence; C, fruit. (b) Circos plot of *W. villosa* genome assembly. The window size 100 kb. A, chromosome karyotypes; B, GC content; C, non-coding RNA (ncRNA) density; D, gene density; E, repeat sequence densities shown as the distribution densities from high (red) to low (green); F, long terminal repeat (LTR) densities shown as the distribution densities from high (red) to low (green); G, syntenic blocks.

genes, respectively (Table S5). Third, the *W. villosa* genome has an LTR Assembly Index (LAI) score of 16.18, falling in the category of 'reference' quality. Fourth, 98.25% of the Illumina and 99.99% of the ONT genome sequencing reads can be mapped back to the *W. villosa* genome. In addition, we generated 163.46 Gb (546.18 million pairs of 150 bp read) of Illumina RNA-seq data and 38 322 PacBio Iso-Seq transcripts, 96.93 and 99.40% of which can be mapped back to the genome, respectively (for the read number and mapping rate of each organ-specific dataset, see Table S6). Finally, the scatter plot of GC depth and GC content showed no indication of contamination in the data (Figure S3). In summary, these quality control metrics all indicate that the *W. villosa* genome assembly is complete and reliable.

Our genome annotation identified 42 588 protein-coding genes in *W. villosa*, with an average gene length of 5277 bp and an average coding sequence (CDS) length of 1192 bp (Table 1 and Table S7). BUSCO evaluation of the annotated proteome of *W. villosa* revealed a high completeness score of 95.2%, closely matching that of the genome. Of the 42 588 annotated proteins, 40 261 (94.5%) showed significant similarity to known sequences in the UniRef90 database, thus being supported by homology evidence. Furthermore, most proteins have functional annotations from at least one of the following sources, including InterPro domains (https://www.ebi.ac.uk/interpro/) (84.0%), COG (www.ncbi.nlm.nih.gov/COG) categories (63.8%), Gene Ontology (GO) (http://geneontology.org) terms (44.9%), and Kyoto Encyclopedia of Genes and Genomes (KEGG) (www.genome.jp/kegg) pathways (28.5%). We also annotated 1843 tRNA, 522 rRNA, 162 miRNA, and 6718 snRNA in the *W. villosa* genome (Table S8).

The genome of *W. villosa* (genome size: approximately 2.8 Gb) is substantially larger than other sequenced genomes in Zingiberales, such as *Musa acuminata* (genome size: approximately 500 Mb) and *Zingiber officinale* (genome size: approximately 1.5 Gb), which can mostly be attributed to differences in their repeat contents (Table S7). Our analysis showed that a total of 2518.16 Mb (87.23%) in the *W. villosa* genome was identified as repetitive sequences. Consistent with the patterns in many other plant genomes, long terminal repeats (LTRs) are the most abundant class of transposable element (TE), accounting for 78.26% of *W. villosa* the genome (Figure 1b and Table S9).

Genome evolution

To study the evolution of the *W. villosa* genome, we conducted a comparative genomic analysis of *W. villosa* and nine other monocot and dicot plants, including *Arabidopsis thaliana*, *Carica papaya*, *M. acuminata*, *Medicago truncatula*, *Oryza sativa*, *Populus trichocarpa*, *Sorghum bicolor*, *Vitis vinifera*, and *Z. officinale*. Reconstruction of orthologous gene clusters identified a set of 799 single-copy orthologous genes shared by all 10 plants. In *W. villosa*, most annotated genes were clustered with genes from at least one other species, while 3299 genes were found to be unique to *W. villosa* (Figure 2a,b and Table S10). Notably, approximately 80% (34 062 out of 42 588) of the *W. villosa* genes have homologs in *Z. officinale*.

A phylogenetic tree was constructed using the 799 single-copy orthologs, and the resulting topology was in agreement with the current understanding of the relationships among the 10 species. In particular, *W. villosa* is sister to *Z. officinale*, the type species of Zingiberaceae, and the next closest relative is *M. acuminaata*, which also belongs to Zingiberales (Figure 2c). Furthermore, by using known divergence times between monocots-eudicots, Zingiberales–Poales, *Oryza*–*Sorghum*, and *Arabidopsis*–*Carica* as calibration points, we inferred that the ancestors of *W. villosa* and *Z. officinale* separated approximately 11 million years ago (MYA), whereas the divergence between Zingiberaceae and Musaceae occurred approoximately 60 MYA (Figure 2c). Analysis of gene family evolution showed that 1531 and 731 gene families exhibited significant expansion and contraction, respectively, in the common ancestor of *W. villosa* and *Z. officinale*. At the same time, in the lineage leading to *W. villosa*, 2125 and 1019 gene families experienced significant expansion and contraction, respectively (Figure 2c). In addition, we conducted KEGG enrichment analysis on these expanded gene families to investigate the overrepresentation of metabolic pathways. Interestingly, we found a significant enrichment of pathways associated with secondary metabolite biosynthesis ($q < 0.05$), including 'terpenoid backbone biosynthesis' and 'sesquiterpenoid and triterpenoid biosynthesis', which may be related to the biosynthesis of volatile terpenoids (Figure S4). We also performed GO enrichment analysis of the expanded gene families and found the enrichment of similar functional terms (e.g. 'terpene synthase activity') (Figure S5). Notably, these results provide a valuable resource for understanding the biosynthesis of active ingredients of *W. villosa*.

As mentioned above, the genome of *W. villosa* contains a very high percentage of repeat sequences, particularly LTRs. Therefore, we examined the insertion time of LTRs in the genomes of *W. villosa*, *Z. officinale*, *M. acuminaata*, and *O. sativa*. As a result, a recent burst of LTRs was observed in all four plants (Figure 2d). In comparison, the distribution of LTR insertion time is relatively broader and more flattened in *W. villosa*, suggesting that *W. villosa* has experienced a more extended period of LTR accumulation. Furthermore, to estimate the potential whole-genome duplication (WGD) events in the evolutionary history of *W. villosa*, we performed pairwise comparisons between *W. villosa*, *Z. officinale*, and *M. acuminaata*, as well as self-comparisons of the three genomes, and examined the distributions of synonymous substitution rates (*Ks*) and four-fold synonymous third-codon transversion rates (4DTv) between syntenic genes (Figure 2e and Figure S6). The results showed that a recent WGD event likely occurred in the common ancestor of *W. villosa* and *Z. officinale*, and there was an independent WGD event during the evolution of *M. acuminaata* (Figure 2e).

Identification of genes related to terpenoid biosynthesis

W. villosa is widely used in Chinese medicine and cuisine for its rich production of volatile terpenoids across the

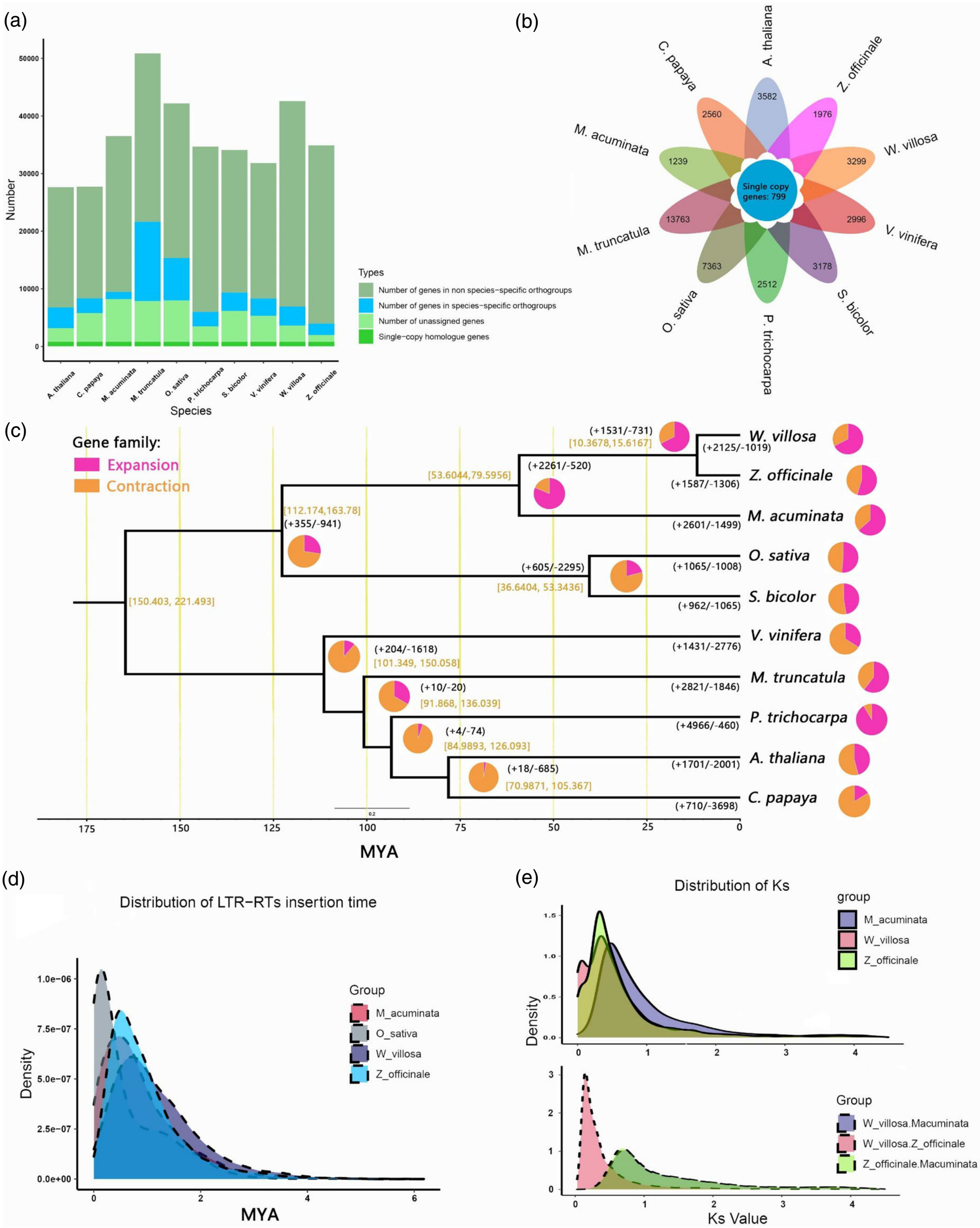

Figure 2. Evolution of the *W. villosa* genome and gene families. (a) Copy number distribution of the gene families in 10 species. (b) Petal diagram of the gene families in 10 species. The middle circle is the number of single-copy orthologous genes shared by all species, and the number of genes in species-specific gene families is on the side. (c) Phylogenetic tree and gene family expansions/contractions in 10 species. (d) Distribution of LTR-RTs insertion time of *W. villosa* and other three plant species. (e) Distribution of *Ks* values between *W. villosa, Z. officinale* and *M. acuminata*. The *Ks* distribution curve of '*W. villosa* vs. *M. acuminata*' is very close to that of '*Z. officinale* and *M. acuminata*', such they mostly overlap with each other.

whole plant. Therefore, we analyzed the volatile terpenoids in seven different organs using GC-MS and detected 25 monoterpenoids, 23 sesquiterpenoids, and two diterpenoids (Figure 3a, Table S11 and Appendix S1). Although diterpenoids were predominantly enriched in leaves and flowers, more than 50% of the total monoterpenoids and sesquiterpenoids were found in seeds (Figure 3b). Overall, 60-DAF (days after flowering) seeds have the richest repertoire of volatile terpenoids because most monoterpenoids and sesquiterpenoids were highly enriched in this organ.

With the identification of volatile terpenoids in *W. villosa*, we next analyzed the genes in relevant biosynthesis pathways. In green plants, precursor molecules for terpenoid biosynthesis are derived from the cytosolic

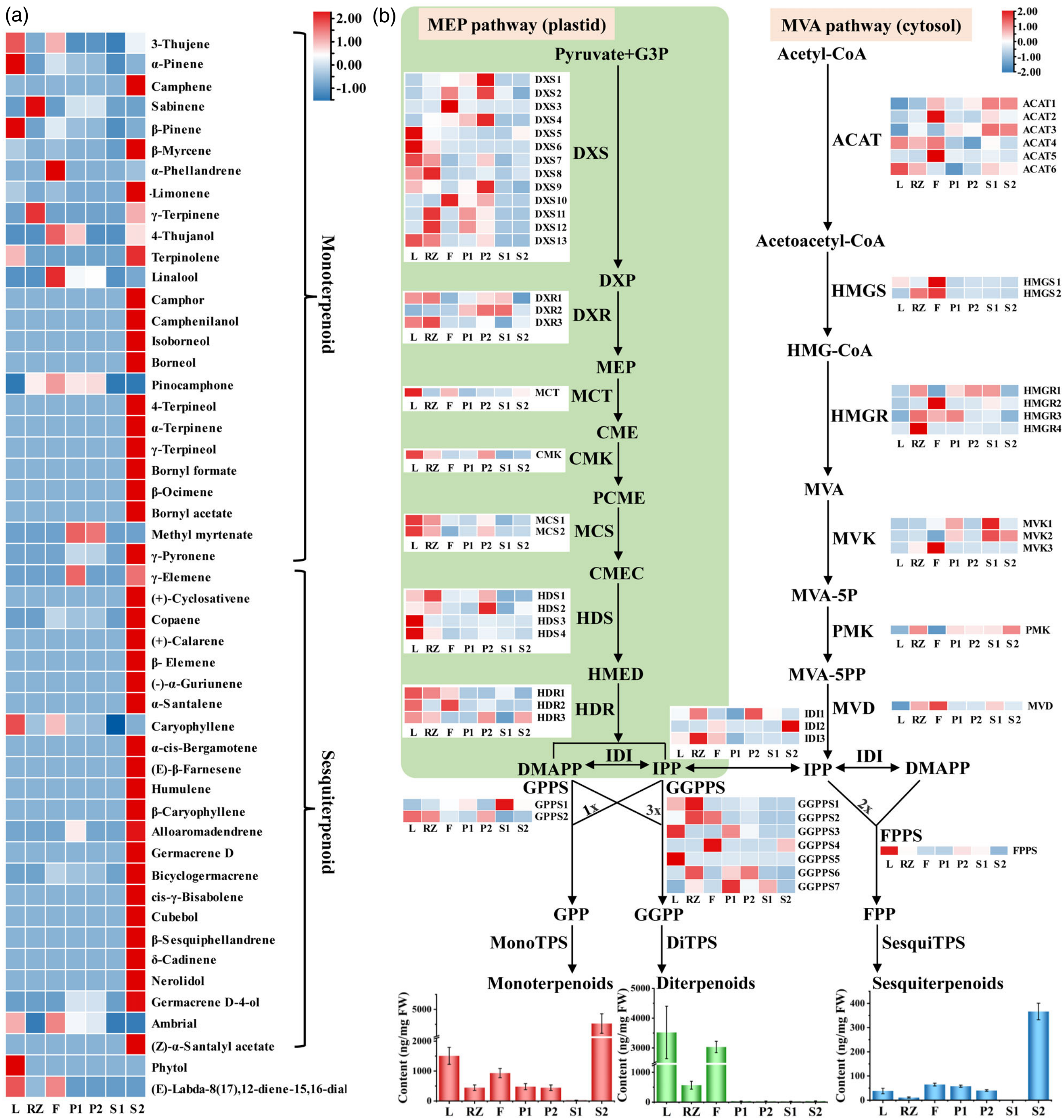

Figure 3. Volatile terpenoids and terpenoid backbone biosynthesis pathways in *W. villosa*. (a) Content heatmap (row scale) of the volatile terpenoids of 50 in seven different organs (L, leaf; RZ, rhizome; F, flower; P1, pericarp of 30-DAF (days after flowering) fruit; P2, pericarp of 60-DAF fruit; S1, seed of 30-DAF fruit; S2, seed of 60-DAF fruit). (b) Tissue-specific expression profiles of genes implicated in terpenoid backbone biosynthesis (heatmap, row scale). The red, green and blue histograms indicate the content of monoterpenoids, diterpenoids and sesquiterpenoids, respectively.

mevalonate and plastidial 2-C-methyl-D-erythritol-4-phosphate pathways (Vranova et al., 2013). Here, the genes involved in terpenoid backbone biosynthesis were identified and compared with their homologs in 15 other plants as reported by Tu et al. (2020). The results showed that the copy numbers of genes encoding 1-deoxy-D-xylulose-5-phosphate synthase and geranyl diphosphate synthase, which may be the rate-limiting enzymes in monoterpenoid biosynthesis, were expanded in *W. villosa* (Figure 3b). By examining their expression profiles in seven different organs, we found that few genes in the terpenoid backbone biosynthesis pathway were specifically highly expressed in 60-DAF seeds (Figure 3b and Table S12). Therefore, we speculate that genes downstream in the volatile terpenoid biosynthesis pathway (e.g. TPSs) might be responsible for the accumulation of monoterpenoids and sesquiterpenoids in 60-DAF seeds.

TPSs are rate-limiting enzymes and use geranyl diphosphate (GPP), farnesyl diphosphate (FPP), and geranylgeranyl diphosphate (GGPP) as direct precursors to synthesize monoterpenes, sesquiterpenes, diterpenes, and triterpenes. Strikingly, we identified 66 putative *WvTPSs* in the genome of *W. villosa* (Table S13), considerably more than the numbers of TPSs reported in related plant genomes. Based on a phylogenetic analysis of 208 TPSs from five representative species (Figure 4a), we classified the 66 *WvTPSs* into five previously recognized TPS subfamilies: TPS-a (24), TPS-b (26), TPS-c (4), TPS-e/f (6), and TPS-g (6). Notably, TPS-a and TPS-b subfamilies are significantly expanded in *W. villosa* compared to other plants (Chen et al., 2011), presumably contributing to the mass production of volatile monoterpenoids and sesquiterpenoids. We also compared the expression profiles of *WvTPSs* in seven different organs (Figure 4b and Table S14) and found that a total of 25 *WvTPSs* (including 20 genes belonging to TPS-a and TPS-b subfamilies) exhibited higher transcript abundance in fruits, suggesting that these genes might have a critical role in the biosynthesis of volatile terpenoids in *W. villosa* fruits.

The 66 *WvTPSs* were distributed on eight chromosomes and several unanchored contigs, among which 44 genes (66.7%) were located in tandem gene clusters, suggesting that tandem duplication events have an important contribution to the expansion of *WvTPSs* (Figure 4c). In total, there are 13 tandem gene clusters with sizes ranging from two to six; genes in one cluster are recent duplicates in the same subfamily and possess a similar exon–intron structure in general (Figure 4c and Figure S7). Tissue-specific transcriptome analysis showed that genes in cluster 3 (*WvTPS11* and *WvTPS14*), cluster 7 (*WvTPS27* and *WvTPS28*), and cluster 11 (*WvTPS43*, *44*, *45*, and *47*) were enriched in seeds of 60-DAF fruit, suggesting that these genes may have important roles for terpenoid synthesis in the seeds. In addition, several clusters (e.g. *WvTPS15*/*WvTPS16*/*WvTPS17* in cluster 4 and *WvTPS42*/*WvTPS46* in cluster 11) exhibited differential expression patterns among tandem duplicates, suggesting rapid divergence in their regulation after gene duplication. We also identified 10 conserved WvTPS protein motifs using MEME (https://meme-suite.org/meme/doc/meme.html), and their lengths ranged from 15 to 41 amino acids (Table S15 and Figure S8). Despite the different types of motifs among some branches, WvTPSs within the same branch generally possessed similar motifs.

Functional characterization of *WvTPSs*

To further investigate the members of WvTPS involved in terpenoid biosynthesis, 18 WvTPSs were cloned and functionally characterized *in vitro* (Appendix S1), including three previously studied enzymes, namely WvTPS14 (previously named AvTPS3, a bornyl diphosphate synthase, or BPPS), WvTPS37 (previously named AvTPS2, a linalool synthase), and WvTPS63 (previously named AvTPS1, a pinene synthase) (Wang et al., 2018; Zhao et al., 2021). Furthermore, as the most important TPS of *W. villosa*, the recombinant WvTPS14 with N-terminal 47 amino acid residues truncated was also analyzed. GC-MS analyses of the catalytic products of the 18 recombinant WvTPS enzymes detected a wide variety of metabolites, including 19 monoterpenoids, 18 sesquiterpenoids, and one diterpenoid (Figure 5a, Tables S16, S17, and Figures S9–S11). These results indicate that most WvTPSs can catalyze the production of multiple products.

Eight WvTPS belonging to the TPS-b subfamily were characterized as monoterpene synthases by their ability to catalyze the production of monoterpenoids from GPP (Figure 5a). Interestingly, WvTPS35 not only catalyzed GPP to form several monoterpenoids, but also catalyzed FPP to produce nerolidol (Figure 5a). Similar to previous results for WvTPS14 (AvTPS3/AvBPPS with N-terminal 26 amino acid residues truncated), the recombinant WvTPS14 constructed in the present study with longer N-terminal truncation (47 amino acid residues) had bornyl diphosphate as its major product and several monoterpenoids as its minor products; at the same time, it also produced α-pinene as an extra minor product, which was not detected in the previous work (Wang et al., 2018). We also compared the catalytic products of WvTPS14 with WvTPS11 and WvTPS13, which shared 97 and 93% identity with WvTPS14 at the protein sequence level, respectively. Similar to WvTPS14, both WvTPS11 and WvTPS13 produced limonene as one of their products; however, no borneol was detected in the enzymatic products of WvTPS11 and WvTPS13 after dephosphorylation, indicating that they did not produce bornyl diphosphate (BPP), whereas WvTPS14 did (Table S16). In addition, both our transcriptome and quantitative RT-PCR (qRT-PCR) data indicated that *WvTPS14* was expressed predominantly in seeds, which is consistent

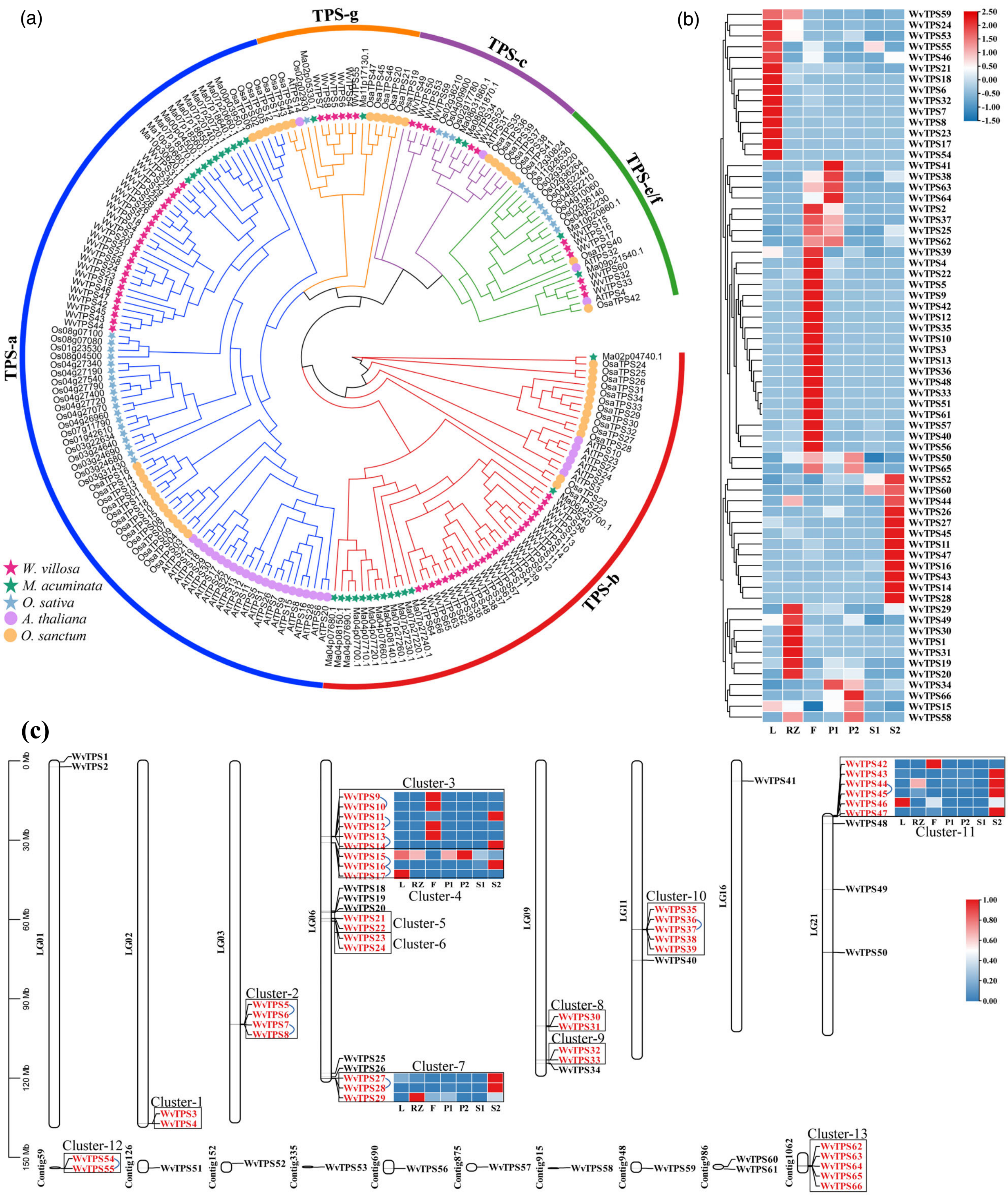

Figure 4. Analysis of TPS gene family in *W. villosa*. (a) Phylogenetic tree of TPS genes from *W. villlosa* (66 genes), *M. acuminata* (31 genes), *O. sativa* (32 genes), *A. thaliana* (32 genes), and *O. sanctum* (47 genes). The outer circle and branch colors represent different TPS gene subclades. Stars represent monocotyledonous plants and circles represent dicotyledonous plants. (b) Heatmap (row scale) showing the differential expression of *WvTPSs* according to the transcriptome data from various organs (L, leaf; RZ, rhizome; F, flower; P1, pericarp of 30-DAF (days after flowering) fruit; P2, pericarp of 60-DAF fruit; S1, seed of 30-DAF fruit; S2, seed of 60-DAF fruit). (c) Schematic map presentation of the genomic localization of 66 *WvTPSs* and expression profiles of predominantly expressed gene clusters in seeds of 60-DAF. Red fonts indicate tandem gene clusters and blue connecting lines indicate tandem gene pairs.

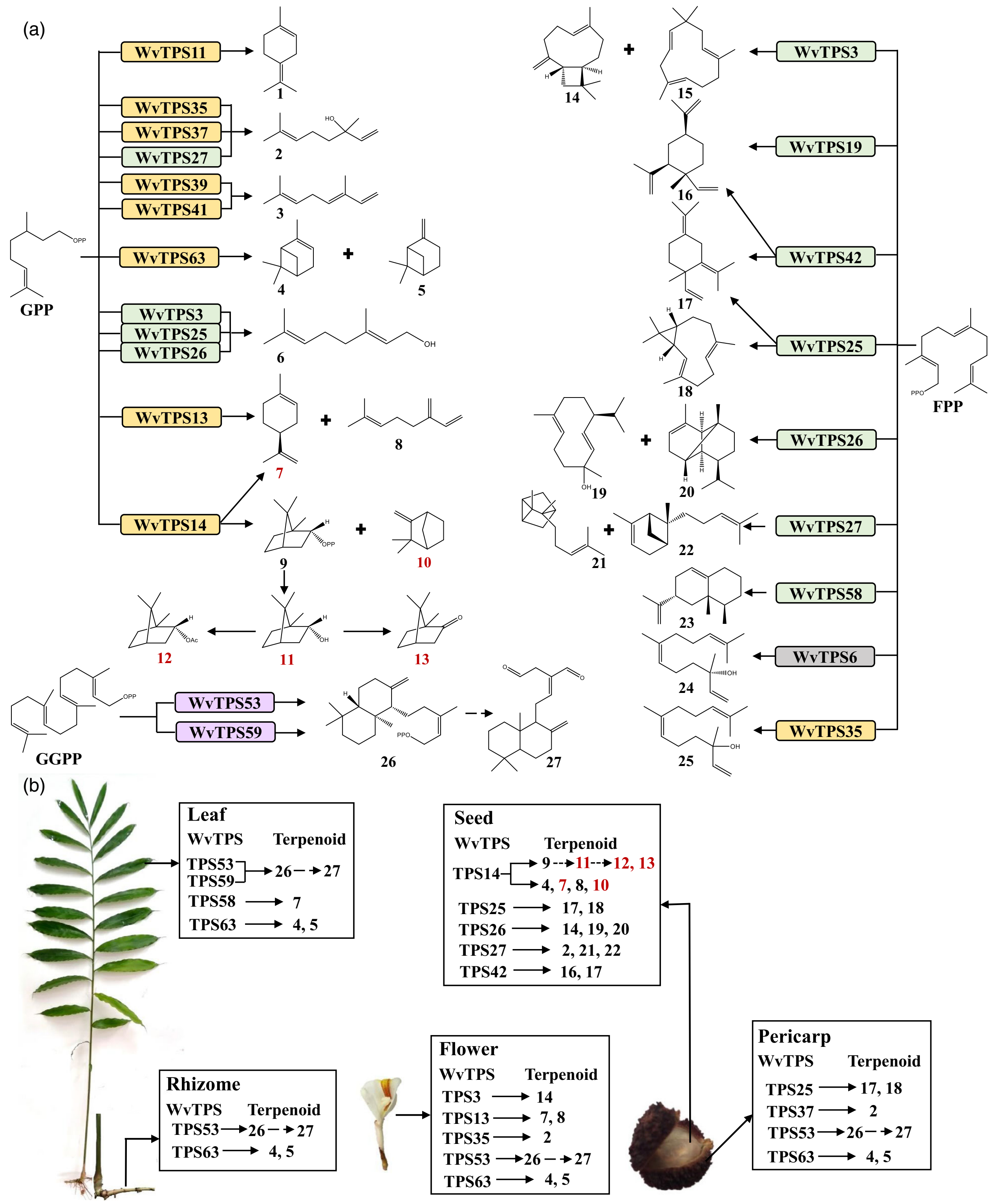

Figure 5. Major products of 18 WvTPS (a) and correlation between WvTPS and tepenoid metablites in five organs of *W. villosa* (b). Rectangular background colors depict different subgroups of TPS genes: TPS-a (green), TPS-b (yellow), TPS-c (purple), and TPS-g (gray). The major terpenoids in seeds were marked in red. GPP, geranyl diphosphate; FPP, farnesyl diphosphate; GGPP, geranylgeranyl diphosphate. 1, terpinolene; 2, linalool; 3, β-ocimene; 4, α-pinene; 5, β-pinene; 6, geraniol; 7, limonene; 8, β-myrcene; 9, bornyl diphosphate; 10, camphene; 11, borneol; 12, bornyl acetate; 13, camphor; 14, caryophyllene; 15, humulene; 16, β-elemene; 17, γ-elemene; 18, bicyclogermacrene; 19, D-germacren-4-ol; 20, copaene; 21, α-santalene; 22, α-cis-bergamotene; 23, aristolochene; 24, *trans*-nerolidol; 25, nerolidol; 26, copalyl diphosphate; 27, (*E*)-labda-8(17),12-diene-15,16-dial.

with the high contents of borneol, bornyl acetate, camphor (three BPP-related terpenoids), and camphene, whereas *WvTPS13* was mainly expressed in flowers (Tables S14, S18, Figure S12 and Figure 3a, 4b).

Most WvTPSs in the TPS-a subfamily were bifunctional because they catalyzed GPP to form monoterpenoids and FPP to form sesquiterpenoids, except for WvTPS42, which only produced sesquiterpenoids (Figure 5a, Tables S16 and S17). Interestingly, when GPP was used as the only substrate, WvTPS3, WvTPS25, and WvTPS26 produced geraniol as their main monoterpenoid product. However, when the substrate was a mixture of GPP and FPP in equal proportion, WvTPS58 only produced sesquiterpenoids, whereas WvTPS25 and WvTPS26 still produced both monoterpenoids and sesquiterpenoids, with sesquiterpenoids becoming the main product (Table S19). Therefore, we speculate that these bifunctional TPSs have higher substrate-selectivity on FPP than on GPP, consistent with the assumed function of the TPS-a subfamily as sesquiterpenoid synthases. Among these bifunctional WvTPSs, our transcriptome and qRT-PCR data indicated that *WvTPS26* was expressed predominantly in seeds. In addition, functional data indicated that WvTPS26 had the highest product diversity. These results indicate that WvTPS26 might play an important role in the synthesis of volatile terpenoids in seeds.

WvTPS6 lacks the RKX_8W motif and belongs to the TPS-g subfamily, and it only catalyzed FPP to form *trans*-nerolidol. WvTPS53 and WvTPS59 belong to the TPS-c subfamily and catalyzed GGPP to produce copalyl diphosphate, the dephosphorized product of which is copalol (Figure S11). Therefore, WvTPS53 and WvTPS59 were characterized as copalyl diphosphate synthase, a class II diTPS.

To identify key TPSs responsible for terpenoid synthesis in *W. villosa* seeds, we performed a correlation analysis between the terpenoid content of seeds and *in vitro* activities of WvTPSs. We found that *WvTPS11*, *WvTPS14*, *WvTPS26*, and *WvTPS27*, which were preferentially expressed in seeds, were positively correlated with bornyl acetate, camphor, borneol, camphene, and limonene (the main terpenoids accumulated in seeds), suggesting that these four genes may be primarily responsible for terpenoid synthesis in the seeds (Figure S13). In addition, the catalytic products and expression patterns of these WvTPSs in five organs are summarized in Figure 5b, revealing that seeds have the greatest number of organ-specific *WvTPSs*, especially *WvTPS14* and *WvTPS26* encoding multi-product TPSs. These results provide an important genetic basis for the high abundance of volatile terpenoids (including bornyl acetate and borneol) in *W. villosa* seeds.

DISCUSSION

The Zingiberaceae family includes approximately 1500 species, most of which have economic value for medicinal, culinary, and ornamental purposes. However, besides a few chloroplast genomes, the whole-genome sequence has only been available for *Z. officinale* so far (Cheng et al., 2021; Li et al., 2021). *W. villosa* is an important medicinal and edible traditional plant, and its genomic information will be valuable for investigating the mechanisms of biosynthesis and accumulation of pharmacodynamic components and the evolution of the Zingiberaceae. Here, we report a high-quality chromosome-level genome assembly of *W. villosa*, with a contig N50 of 9.13 Mb, higher than that of other medicinal plants (Yang et al., 2021, 2022), such as *Z. officinale* (Li et al., 2021) (1.53 Gb, N50 of 4.68 Mb). In total, 42 588 genes were annotated in *W. villosa* genome, which is more than *Z. officinale* (36 503 genes) and *M. acuminata* (36 542 genes) (D'Hont et al., 2012). In total, 2125 gene families experienced significant expansion in the lineage leading to *W. villosa*. GO terms related to terpenoid metabolism, such as 'terpene synthase activity', were significantly enriched in these gene families, suggesting that *W. villosa* has accumulated genes involved in terpenoid synthesis during its recent evolution. *Ks* and 4DTv analyses both suggest a recent WGD event in the evolutionary history of *W. villosa*, which likely coins the recent WGD reported in *Z. officinale* and thus might be shared by other Zingiberaceae as well (Cheng et al., 2021; Li et al., 2021). This WGD event may have contributed to the species evolution, genome size variation, chromosomal rearrangement, and gene family expansion/contraction in Zingiberaceae.

The main effective components of *W. villosa* fruits are the volatile terpenoids comprising a rich array of bornyl acetate, camphor, limonene, camphene, and borneol. Analyses of the *TPS* gene family in *Eucalyptus grandis* and *Cinnamomum kanehira* suggest that significant expansions of *TPS-a* and *TPS-b* subfamilies may have contributed to the biosynthesis and diversity of volatile monoterpenoids and sesquiterpenoids, and that polyploidization may further influence the evolution of terpenoid metabolism (Chaw et al., 2019; Myburg et al., 2014). In the present study, we also observed a considerable expansion of the *TPS* gene family in *W. villosa*, especially in the *TPS-a* and *TPS-b* subfamilies. Tissue-specific transcriptome analysis revealed 12 *WvTPS* genes that were predominantly expressed in the seeds of 60-DAF fruits, which is consistent with the enrichment of volatile terpenoids in that tissue and the enzymatic activity of the genes. These results suggest that these genes may have important roles in the content and diversity of volatile terpenoids in the seeds of *W. villosa*.

Previous studies have shown that tandem duplication is a major evolutionary force driving the expansion of the *TPS* gene family in various plants (Chen et al., 2020a,b,c). In the present study, the analyses of enzymes involved in terpenoid biosynthesis pathways suggest that the expansion of the *TPS* family, primarily through tandem gene duplications, may have led to the terpenoid diversity in *W. villosa*. It has been reported that the catalytic production of TPS is related to the sequence similarity (Karunanithi et al., 2020; Wang et al., 2021); in general, WvTPSs in the same tandem gene cluster tend to have closer evolutionary relationships and more similar catalytic products. For example, WvTPS35 and WvTPS37 from tandem gene cluster 10 were found to have similar catalytic products (Figure 5a; Figure S14). Interestingly, however, in cluster 3, the duplicated gene pair WvTPS13 and WvTPS14 demonstrated highly similar sequences but different functions; only WvTPS14 can catalyze the formation of BPP, the important precursor of bornyl acetate, representing a potential case of neo-functionalization. Furthermore, we found that *WvTPS14*, which has a tissue-specific high expression in seeds (TPM = 3543.4), may be mainly responsible for the synthesis of pharmacodynamic terpenoids in seeds, whereas *WvTPS13*, which is highly expressed in flowers (TPM = 8049.7), probably plays major roles in the synthesis of floral fragrance and defense response-related terpenoids. These results suggest that duplication of *TPS* genes and the subsequent sub- (or neo-) functionalization may facilitate the segregation of biological properties and that the high gene expression of genes may promote the production of the main components of terpenoids.

Terpenoids are the main active ingredients underlying the excellent edible and medicinal values of *W. villosa*, although the molecular mechanism of their biosynthesis remains largely unknown. Here, we performed a functional characterization of 18 WvTPSs, focusing on the TPS-a and TPS-b subfamilies, which mostly catalyze the production of monoterpenoids and sesquiterpenoids. In total, 14 WvTPSs were characterized as monoterpenoid or sesquiterpenoid synthases producing muti-products in the present study. Remarkably, WvTPS14 and WvTPS26, both encoding multi-product enzymes, are mainly expressed in the 60-DAF seeds, which have a rich diversity of volatile terpenoids. Importantly, WvTPS14 was found to be responsible for the synthesis of borneol-related terpenes, which accounted for 66.5% of the total volatile terpenoids content in 60-DAF seeds. According to the literature, only one BPPS has been found in each of *Cinnamomum burmannii*, *Lavandula angustifolia*, and *Salvia officinalis* so far (Despinasse et al., 2017; Ma et al., 2022; Radwan et al., 2017). In the present study, we also found that WvTPS11 and WvTPS13, the two closest relatives of WvTPS14, do not function as BPPS, although they have other catalytic products in common, such as limonene and β-myrcene. In addition, the other five enzymes in the TPS-b subfamily (monoterpene synthase) do not function as BPPS either, and their catalytic products differ significantly from that of WvTPS14, WvTPS11, and WvTPS13. Meanwhile, *WvTPS14* was the only member of the *TPS-b* subfamily that was significantly highly expressed in the seeds. Therefore, our results suggest that WvTPS14 is the only BPPS in *W. villosa*, providing a basis for further dissection of terpenoid biosynthesis and TPS functional diversification in *W. villosa*.

A large number of bifunctional TPSs catalyzing the production of terpenoids have previously been identified in *Cannabis sativa* and *Setaria italica* (Booth et al., 2020; Karunanithi et al., 2020). In the present study, seven WvTPSs were identified as bifunctional TPSs *in vitro*, including six TPS-a subfamily members and one TPS-b member; five bifunctional WvTPSs catalyzed the transformation of GPP to geraniol *in vitro*; however, geraniol was not detected in *W. villosa*, which might be related to the protein subcellular location and the endogenous substrate availability. β-ocimene is a signal molecule involved in plant defense (Faldt et al., 2003). Both WvTPS39 and WvTPS41 produce β-ocimene as the main product, although they lack the conserved NSE/DTE motif present in other β-ocimene synthases (Figure S15). To our knowledge, this represents the first report of β-ocimene synthases lacking the NSE/DTE motif, which is the binding region of metal ions. The relationship between the presence/absence of conserved motifs and the catalytic activities of TPSs is worthy of further studies (Karunanithi & Zerbe, 2019).

CONCLUSIONS

The present study reports a high-quality chromosome-level reference genome of *W. villosa* with comprehensive genomic, transcriptomic, and metabolic analyses, as well as the identification and functional characterization of TPS-encoding genes, which can provide insights into the molecular genetic basis of the diversity and abundance of volatile terpenoids. Importantly, 66 *WvTPS* genes were identified in the *W. villosa* genome, among which 18 were functionally characterized, and most of the enzymes were found to be product diverse. Therefore, we consider that our genome data will contribute to functional genomic research and genome-assisted breeding for *W. villosa*.

EXPERIMENTAL PROCEDURES

Plant materials

The plant materials of 'Yuanguo', a cultivar of *W. villosa*, were collected from Yangchun, Guangdong Province, China, which is considered the authentic production area of *W. villosa* (Figure 1a). For genomic sequencing, fresh and healthy leaves were harvested. For transcriptome and metabolome analysis, three biological replicates were collected from each of the following seven organs: L, leaf; RZ, rhizome; F, flower; P1, pericarp of 30-DAF; P2, pericarp of 60-DAF, S1, seeds of 30-DAF; S2, seeds of 60-DAF. All collected samples were washed with ultrapure water immediately, frozen in liquid nitrogen, and stored at −80°C.

Oxford Nanopore sequencing library construction and sequencing

Genomic DNA was extracted from the fresh leaves of *W. villosa* using a Genomic DNA extraction kit (catalog. no. 13323; Qiagen, Hilden, Germany). The extracted DNA was assayed for DNA purity using a NanoDrop One UV-Vis spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA), followed by accurate DNA quantification using a Qubit® 3.0 Fluorometer (Invitrogen, Waltham, MA, USA). To construct the Nanopore sequencing library, long DNA fragments were size-selected using the BluePippin system (Sage Science, Beverly, MA, USA). Next, the DNA was repaired and the DNA ends were prepared for adapter attachment. The sequencing adapters provided in the SQK-LSK109 kit were then ligated to the DNA ends and, finally, the size of the library fragments was quantified using the Qubit® 3.0 fluorometer. Then, the sequencing adapters supplied in the SQK-LSK109 kit were attached to the DNA ends and, finally, the size of Library fragments was quantified using the Qubit® 3.0 Fluorometer. Single-molecule real-time (SMRT) sequencing of the purified library was performed on a GridION/PromethION sequencer (Oxford Nanopore Technologies, Oxford, UK) with six flow cells.

Hi-C library construction and sequencing

The Hi-C library was prepared with a improved procedure. Briefly, fresh leaves were fixed with formaldehyde to cross-link DNA to protein and protein to protein. Fixed samples were then lysed, and the chromatin was digested with the restriction endonuclease *Dpn*II. Biotin-labeled bases were introduced during the blunt-end ligation process. The ligated DNA was sheared into 300–600-bp fragments, then blunt-end repaired and A-tailed, followed by amplification to obtain library products. The libraries were sequenced using the Illumina NovaSeq platform (Illumina, San Diego, CA, USA) under paired-end 150 bp mode.

Transcriptome library construction and sequencing

Transcriptome data were generated using two sequencing approaches. For short-read RNA-seq analysis, RNA from seven different tissues was extracted using TRNzol Universal Kit (Tiangen, Beijing, China). The TruSeq RNA Library Preparation Kit (Illumina) was used to generate RNA libraries in accordance with the manufacturer's recommendations, followed by sequencing on the Illumina NovaSeq platform. The resulting short-read sequencing data were aligned to the *W. villosa* genome using HISAT2 (version 2.2.1) (Kim et al., 2019). The transcripts per million (TPM) values were calculated to measure gene expression levels. For long-read RNA-seq analysis, equal concentrations of RNAs from different organs were mixed, and a 20-kb SMRTbell Template library was prepared and sequenced on a PacBio Sequel platform (Pacbio, Menlo Park, CA, USA).

Genome size and heterozygosity estimation

To estimate the genome size of *W. villosa*, a DNA library with an insert size of 400 bp was constructed for sequencing on the Illumina NovaSeq platform. The software JELLYFISH (2.2.9) (Marçais & Kingsford, 2011) was used to calculate the 17-mer frequency distribution, and GENOMESCOPE (version 2.0) (Ranallo-Benavidez et al., 2020) was used to estimate the genome size and heterozygosity rate.

Chromosome-level genome assembly

The filtered Nanopore reads were corrected using NEXTDENOVO (version 1.0; https://github.com/Nextomics/NextDenovo) with the parameters: seed_cutoff = 25 k. Then, the corrected reads were assembled with SMARTDENOVO (https://github.com/ruanjue/smartdenovo) with the parameters: -k 17, −J 3000. To further improve the assembly accuracy, iterative polishing was performed using NEXTPOLISH (https://github.com/Nextomics/NextPolish), including two rounds of polishing using the Nanopore long reads and two rounds of polishing using the Illumina short reads.

A chromosome-level assembly was constructed from the draft contig-level assembly. First, quality controlling of raw Hi-C data was performed using HI-C-PRO (version 2.8.0) (Servant et al., 2015); then, FASTP (version 0.12.6) (Chen et al., 2018a, b) was used to filter out low-quality sequences (quality scores < 20), adaptor sequences, and sequences shorter than 30 bp; BOWTIE2, version 2.3.2 (Langmead & Salzberg, 2012) was used to map clean paired-end reads to the draft assembly; finally, LACHESIS (ligating adjacent chromatin enables scaffolding *in situ*) (Burton et al., 2013) was used to produce chromosome-level scaffolds.

Genome assembly quality assessment

Genome assembly accuracy and completeness were first assessed using the Hi-C interaction heatmap. Second, GC depth scatter plots were used to assess the presence of contamination in the sequencing data. Third, the second-generation and third-generation genome sequencing data were mapped to the genome using BWA (version 0.7.12) (Li & Durbin, 2009) and MINIMAP2 (version 2.17) (Li H, Li, 2018), respectively, to assess their coverage. In addition, the second-generation transcriptome sequencing data were mapped to the genome using HISAT2 (version 2.2.1) (Kim et al., 2019), whereas the PacBio Iso-Seq data were processed by ISOSEQ3 (version 3.4.0) (https://github.com/PacificBiosciences/IsoSeq) and PBMM2 (version 1.7.0) (https://github.com/PacificBiosciences/pbmm2). Finally, the assembled genome was also analyzed using BUSCO (version 5.2.2) (Waterhouse et al., 2018) and LTR_RETRIEVER (version 2.9.0) (Ou et al., 2018) to evaluate the completeness and accuracy of the genome.

Genome annotation

Homology-based and *de novo* approaches were applied to identify TE in the *W. villosa* genome. Briefly, a *de novo* repeat library of *W. villosa* genome was constructed using REPEATMODELER (version 2.0.1) (http://www.repeatmasker.org/RepeatModeler) with the '-LTRStruct' option. The obtained library was then combined with known repeats of Zingiberales in the Repbase (http://www.girinst.org/repbase) database to identify repetitive sequences in the *W. villosa* genome using REPEATMASKER (version 4.1.1) (http://www.repeatmasker.org/).

miRNA, rRNA, and snRNA genes were detected using INFERNAL (version 1.1.2) (Nawrocki et al., 2009) to search the Rfam (Gardner et al., 2009) database with the default parameters. tRNAs were predicted using TRNASCAN-SE (version 2.0.9) (Lowe & Chan, 2016).

Protein-coding genes were annotated using a combination of homology-evidence, RNA-seq data, and *ab initio* gene prediction methods. In brief, homology-based gene models were first generated using EXONERATE (version 2.2.0) (Slater & Birney, 2005) based on all monocot protein sequences in the OrthoDB database, and the models were used to train three ab initio predictors: AUGUSTUS (version 3.4.0) (Testa et al., 2015), GENEMARK-ES (version 4.58) (Borodovsky & Lomsadze, 2011), and SNAP, version 2013-11-29 (Bischoff & Schmidt, 2006). At the same time, a *de novo* transcriptome assembly was generated using TRINITY (version 2.9.0) (Haas et al., 2013). The trained predictors, homology-evidence, and transcriptome assembly were used as input of the MAKER (version 2.31) pipeline (Campbell et al., 2014) to conduct genome annotation. The resulting gene predictions were then polished using PASA (version 2.4.1) (Haas et al., 2008). The longest transcript was retained for each gene model.

Functional annotation (e.g. COG categories, GO terms, and KEGG pathways) of the protein-coding genes was carried out using EGGNOG-MAPPER (version 2.0.5) (Huerta-Cepas et al., 2017), and protein domains were identified using INTERPROSCAN (5.47-82.0) (Jones et al., 2014). Finally, two strategies were taken to evaluate the accuracy of genome annotation. First, the RNA-seq data from the seven different tissues were mapped to the coding sequences with BOWTIE2 to calculate the transcriptome support rate. Second, BUSCO was used to assess the completeness of gene annotations.

Genome evolution

To investigate evolutionary relationships of *W. villosa*, its predicted proteomes and that of nine other plants, including *Z. officinale*, *M. acuminata*, *O. sativa*, *S. bicolor*, *V. vinifera*, *M. truncatula*, *P. trichocarpa*, *A. thaliana*, and *C. papaya*, were used to construct orthologous groups using ORTHOFINDER (version 2.5.1) (Li et al., 2003). The protein sequences of single-copy orthologs genes from 10 species were used for the phylogenetic reconstruction. MAFFT (version 7.471) (Katoh & Standley, 2013) was used to align the protein sequences, and then poorly aligned regions were trimmed using GBLOCKS (version 0.91b) (Gastresana, 2000). The maximum-likelihood tree was constructed using RAXML (Stamatakis, 2006) with 1000 bootstrap replicates and visualized using FIGTREE (https://github.com/rambaut/figtree). MCMCTREE from the PAML packages (version 4.9i) (Yang, 2007) was used to calculate divergence times. Four calibration points were obtained from the TimeTree database (http://www.timetree.org): monocots–eudicots (115–308 Mya), Zingiberales–Poales (97–116 Mya), *Oryza*–*Sorghum* (42–52 Mya), and *Arabidopsis*–*Carica* (63–82 Mya). Analysis of gene family evolution was then conducted using CAFE (version 4.3) (De et al., 2006), which uses a birth and death process to model gene gain and loss over a phylogeny. The *Ks* and 4DTv values were calculated to infer the occurrence of WGD events. WGDI (version 0.4.9) was used to identify gene pairs in collinear intervals for *Ks* calculation (Wang et al., 2010). 4DTv values were calculated using the Perl script calculate_4DTV_correction.pl (https://github.com/JinfengChen/Scripts).

Identification of genes related to terpenoid biosynthesis

To investigate the genes involved in the terpenoid skeleton synthesis pathways, we first retrieved protein sequences from *A. thaliana* genome, including 1-deoxy- D -xylulose-5-phosphate synthase (*DXS*), 1-deoxy- D -xylulose-5-phosphate reductoisomerase (*DXR*), 2-C-methyl- D -erythritol 4-phosphate cytidylyltransferase (*MCT*), 4-diphosphocytidyl-2- C -methyl-D-erythritol kinase (*CMK*), 2-C-methyl- D -erythritol-2,4-cyclodiphosphate synthase (*MCS*), (*E*)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase (*HDS*), 4-hydroxy-3-methylbut-2-enyl-diphosphate reductase (*HDR*), acyl-coenzyme A-cholesterol acyltransferase (*ACAT*), hydroxymethylglutaryl-CoA synthase (*HMGS*), hydroxymethylglutaryl-CoA reductase (*HMGR*), mevalonate kinase (*MVK*), phosphomevalonate kinase (*PMK*), mevalonate diphosphate decarboxylase (*MVD*), isopentenyl-diphosphate isomerase (*IDI*), geranyl diphosphate synthase (*GPPS*), geranylgeranyl diphosphate synthase (*GGPPS*), and farnesyl diphosphate synthase (*FPPS*) from the NCBI database (https://www.ncbi.nlm.nih.gov). Their homologs in the genomes of *W. villosa* were identified using iterative BLASTP (https://blast.ncbi.nlm.nih.gov/Blast.cgi) searches with an *E*-value cutoff of 1×10^{-5}; at each iteration, newly discovered homologs were used as queries to carry out the next iteration of BLASTP search until no additional homolog can be identified. In addition, for GPPS, GGPPS, and FPPS, candidate genes were further identified by functional annotation and BLASTP identity of > 50%.

To identify the candidate TPS genes, HMM profiles of Terpene_synth (PF01397) and Terpene_synth_C (PF03936) obtained from the Pfam database (http://pfam.xfam.org) were used to search against *W. villosa* protein sequences using HMMER (Johnson et al., 2010) with an *E*-value cutoff of 1×10^{-5}. To classify the TPS genes into different subfamilies, we downloaded the protein sequences of TPS genes from four different plants, including *M. acuminata*, *O. sativa*, *A. thaliana*, *Ocimum sanctum* (Kumar et al., 2018). Then, TPS protein sequences of the five species were aligned using MAFFT, and the maximum-likelihood tree was constructed using IQ-TREE (Nguyen et al., 2015) with 1000 ultra-fast bootstrap replicates. The phylogenetic tree was visualized using EVOLVIEW (Zhang et al., 2012). MEME was used to identify conserved motifs. Analysis and visualization of domain composition, gene structure, and chromosome distribution of TPS genes were conducted using TBTOOLS (Chen et al., 2020a,b,c).

ACKNOWLEDGEMENTS

This work is financially supported by the National Natural Science Foundation of China (81303163 and 81872954), Key-Area Research and Development Program of Guangdong Province (No.2020B020221001, No.2020B020221002, and No.2020B0202090001).

AUTHOR CONTRIBUTIONS

J-FY, X-FZ, and G-ZH conceived and designed the study. H-YZ, J-SW, Y-YZ, X-JL, JS, ML, H-LL, and Y-WS prepared the materials and conducted the experiments. J-FY, X-FZ, PY, F-PL, and XT analyzed the data and prepared the results. PY, X-FZ, and J-FY wrote the manuscript. D-MM, G-ZH, and R-TZ revised the manuscript. All authors read and approved the final version of the manuscript submitted for publication.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The raw genome and transcriptome sequencing data reported in the present study have been deposited in the National Center for Biotechnology Information (NCBI) database under project number PRJNA796955. The whole-genome assembly has been deposited in NCBI under accession number JAKLTH000000000. Additionally, the gene structure annotations, predicted CDS and protein sequences are available at FigShare (https://doi.org/10.6084/m9.figshare.19200005.v1).

SUPPORTING INFORMATION

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

Figure S1. 17-mer analysis to estimate the *W. villosa* genome size.

Figure S2. Heatmap of the Hi-C interaction density between 24 pseudochromosomes. The color from light to dark indicates the increase in the intensity of interaction.

Figure S3. The GC depth distribution of the *W. villosa* genome.

Figure S4. The KEGG pathway analysis of expanded gene families in the *W. villosa* genome.

Figure S5. GO annotations of expanded gene families in the *W. villosa* genome.

Figure S6. Distribution of 4DTv values between *W. villosa, Z. officinale* and *M. acuminata*. The 4DTv distribution curve of '*W. villosa* vs. *M. acuminata*' is very close to that of '*Z. officinale* vs. *M. acuminata*' so they mostly overlap with each other.

Figure S7. Phylogenetic relationships and exon-intron structure of *WvTPSs*. Exon–intron distribution was performed using TBTOOLS software. Orange boxes indicate exons; black lines indicate introns.

Figure S8. Motif structures of *WvTPSs*. Ten classical motifs in WvTPSs were analyzed using the MEME tool. Different color blocks represent different motifs.

Figure S9. The GC-MS chromatograms of products generated by recombinant WvTPSs catalyzing GPP. The red line represents the reaction of recombinant WvTPS with the substrate GPP, and the blue line represents the reaction of boiled recombinant WvTPS (negative control) with GPP. The main product is marked in red.

Figure S10. The GC-MS chromatograms of products generated by recombinant WvTPSs catalyzing FPP. The red line represents the reaction of recombinant WvTPS with the substrate FPP, and the blue line represents the reaction of boiled recombinant WvTPS (negative control) with FPP. The main product is marked in red.

Figure S11. The GC-MS chromatograms of products generated by recombinant WvTPS53 and WvTPS59 catalyzing GGPP. The blue line represents the reaction of recombinant WvTPS with the substrate GGPP, and the red line represents the reaction of boiled recombinant WvTPS (negative control) with GGPP.

Figure S12. The expressional level of *WvTPS* in different organs. The left and right *y*-axis indicates the relative expression level (qRT-PCR) and TPM (transcripts per million) value (transcriptome), respectively. Data represent the mean ± SD ($n = 3$). The pericarp and seed used in this analysis were from 60-DAF fruits.

Figure S13. Correlation analysis between terpenoids from 60-DAF seeds and 18 *WvTPSs* of functional characterization.

Figure S14. Phylogenetic analysis of 18 functionally characterized WvTPSs in *W. villosa*.

Figure S15. Amino acid sequence alignment of *WvTPS39*, *WvTPS41*, *ABY65110.1* and *AMB57287.1*. The red box represents the conserved motifs. *ABY65110.1* from *Phaseolus lunatus* and *AMB57287.1* from *Osmanthus fragrans*.

Figure S16. Mirror MS/MS spectra of the products (red) and standards (blue) used for WvTPS functional characterization in *W. villosa*.

Table S1. Summary of sequencing data for *W. villosa*.

Table S2. Estimation of genome size based on 17-mer statistics.

Table S3. Overview of the genome assembly of *W. villosa*.

Table S4. The contig cluster of 24 pseudochromosomes length.

Table S5. Evaluation of completeness of the final genome assembly and annotation using BUSCO.

Table S6. Percentages of RNA-seq reads mapped to the *W. villosa* genome.

Table S7. Comparison of the gene set of *W. villosa* with other species.

Table S8. Statistics of non-protein-coding gene annotations in the *W. villosa* genome assembly.

Table S9. Repeat annotations of the *W. villosa* genome assembly.

Table S10. Statistics of gene families in 10 plant species.

Table S11. Volatile terpenoids in seven organs of *W. villosa*.

Table S12. List of genes and their expression levels (transcripts per million, TPM) in different organs.

Table S13. *WvTPSs* information for *W. villosa*.

Table S14. List of genes and their expression levels (transcripts per million, TPM) in different organs of *WvTPS*.

Table S15. Distribution of conserved motifs in *W. villosa* TPS proteins based on the results of MEME analysis.

Table S16. Percentages of monoterpenoid products with GPP as substrate.

Table S17. Percentages of sesquiterpenoid products with FPP as substrate.

Table S18. Data of the relative expression level (qRT-PCR) of *WvTPS* in different organs.

Table S19. Percentages of sesquiterpenoid products with GPP and FPP as mixed substrate.

Table S20. Primers used for gene cloning.

Table S21. Primers used for expression vector construction.

Table S22. Primers used for qPCR.

Appendix S1. Detailed methods for GC-MS analysis, gene cloning, prokaryotic expression and protein purification of WvTPS, enzyme assay and product analysis, and quantitative real-time PCR.

REFERENCES

Ao, H., Wang, J., Chen, L., Li, S. & Dai, C. (2019) Comparison of volatile oil between the fruits of *Amomum villosum* Lour. And *Amomum villosum* Lour. Var. xanthioides T. L. Wu et Senjen based on GC-MS and chemometric techniques. *Molecules*, **24**, 1663.

Bischoff, P. & Schmidt, G. (2006) Monitoring methods: SNAP. *Best Practice & Research. Clinical Anaesthesiology*, **20**, 141–146.

Booth, J.K., Yuen, M.M.S., Jancsik, S., Madilao, L.L., Page, J.E. & Bohlmann, J. (2020) Terpenoid synthases and terpene variation in *Cannabis sativa*. *Plant Physiology*, **184**, 130–147.

Borodovsky, M. & Lomsadze, A. (2011) Eukaryotic gene prediction using GeneMark.Hmm-E and GeneMark-ES. *Current Protocols in Bioinformatics*, **35**, 4.6.1–4.6.10.

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O. & Shendure, J. (2013) Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nature Biotechnology*, **31**, 1119–1125.

Campbell, M.S., Holt, C., Moore, B. & Yandell, M. (2014) Genome annotation and curation using MAKER and MAKER-P. *Current Protocols in Bioinformatics*, **48**, 4.11.1-4.11.39.

Chaw, S.M., Liu, Y.C., Wu, Y.W., Wang, H.Y., Lin, C.Y., Wu, C.S. *et al.* (2019) Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution. *Nature Plants*, **5**, 63–73.

Chen, C., Chen, H., Zhang, Y., Thomas, H.R., Frank, M.H., He, Y. *et al.* (2020a) TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Molecular Plant*, **13**, 1194–1202.

Chen, F., Tholl, D., Bohlmann, J. & Pichersky, E. (2011) The family of terpenoid synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *The Plant Journal*, **66**, 212–229.

Chen, L.X., Lai, Y.F., Zhang, W.X., Cai, J., Hu, H., Wang, Y. *et al.* (2020b) Comparison of volatile compounds in different parts of fresh *Amomum villosum* Lour. From different geographical areas using cryogenic grinding combined HS-SPME-GC-MS. *Chinese Medicine*, **15**, 97.

Chen, S., Zhou, Y., Chen, Y. & Gu, J. (2018a) Fastp: an ultra-fasta all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.

Chen, Y.C., Li, Z., Zhao, Y.X., Gao, M., Wang, J.Y., Liu, K.W. *et al.* (2020c) The *Litsea* genome and the evolution of the laurel family. *Nature Communications*, **11**, 1675.

Chen, Z., Ni, W., Yang, C., Zhang, T., Lu, S., Zhao, R. *et al.* (2018b) Therapeutic effect of *Amomum villosum* on inflammatory bowel disease in rats. *Frontiers in Pharmacology*, **9**, 639.

Chen, Z.Y. & Chen, S.J. (1982) Preliminary report of chromosome numbers on Chinese Zingiberaceae. *Guihaia*, **2**, 153–157.

Cheng, S.P., Jia, K.H., Liu, H., Zhang, R.G., Li, Z.C., Zhou, S.S. *et al.* (2021) Haplotype-resolved genome assembly and allele-specific gene expression in cultivated ginger. *Horticulture Research*, **8**, 188.

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

Despinasse, Y., Fiorucci, S., Antonczak, S., Moja, S., Bony, A., Nicole, F. *et al.* (2017) Bornyl-diphosphate synthase from *Lavandula angustifolia*: a major monoterpene synthase involved in essential oil quality. *Phytochemistry*, **137**, 24–33.

D'Hont, A., Denoeud, F., Aury, L.M., Baurens, F.C., Carreel, F., Garsmeur, O. *et al.* (2012) The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plant. *Nature*, **488**, 213–217.

Faldt, J., Arimura, G., Gershenzon, J., Takabayashi, J. & Bohlmann, J. (2003) Functional identification of AtTPS03 as (E)-β-ocimene synthase: a monoterpenoid synthase catalyzing jasmonate- and wound-induced volatile formation in Arabidopsis thaliana. *Planta*, **216**, 745–751.

Gardner, P.P., Daub, J., Tate, J.G., Nawrocki, E.P., Kolbe, D.L., Lindgreen, S. *et al.* (2009) Rfam: updates to the RNA families database. *Nucleic Acids Research*, **37**, D136–D140.

Gastresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, **17**, 540–552.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J. *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nature Protocols*, **8**, 1494–1512.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J. *et al.* (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biology*, **9**, R7.

Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., von Mering, C. *et al.* (2017) Fast genome-wide functional annotation through orthology assignment by eggnog-mapper. *Molecular Biology and Evolution*, **34**, 2115–2122.

Hugo, D.B., Mark, N., Axel, D.P., Jane, D., Tomas, F., Le, T.T.H. *et al.* (2018) Convergent morphology in Alpinieae (Zingiberaceae): recircumscribing *amomum* as a monophyletic genus. *Taxon*, **67**, 6–36.

Johnson, L.S., Eddy, S.R. & Portugaly, E. (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, **11**, 431.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.

Karunanithi, P.S., Berrios, D.I., Wang, S., Davis, J., Shen, T., Fiehn, O. *et al.* (2020) The foxtail millet (*Setaria italica*) terpenoid synthase gene family. *The Plant Journal*, **103**, 781–800.

Karunanithi, P.S. & Zerbe, P. (2019) Terpenoid synthases as metabolic gatekeepers in the evolution of plant terpenoid chemical diversity. *Frontiers in Plant Science*, **10**, 1166.

Katoh, K. & Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.

Ke, B. & Shi, L. (2012) Exploring clinical efficacy of *fructus Amomi*. *China Journal of Traditional Chinese Medichine and Pharmacy*, **27**, 128–129.

Kim, D., Paggi, J.M., Park, C., Bennett, C. & Salzberg, S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, **37**(8), 907–915.

Kumar, Y., Khan, F., Rastogi, S. & Shasany, A.K. (2018) Genome-wide detection of terpenoid synthase genes in holy basil (Ocimum sanctum L.). *PLoS One*, **13**, e0207097.

Langmead, B. & Salzberg, S.L. (2012) Fasta gapped-read alignment with bowtie 2. *Nature Methods*, **9**, 357–359.

Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.

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

Li, H., Wu, L., Dong, Z., Jiang, Y., Jiang, S., Xing, H. *et al.* (2021) Haplotype-resolved genome of diploid ginger (*Zingiber officinale*) and its unique gingerol biosynthetic pathway. *Horticulture Research*, **8**, 189.

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

Lowe, T.M. & Chan, P.P. (2016) tRNAsan-SE on-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Research*, **44**, W54–W57.

Ma, Q., Ma, R., Su, P., Jin, B., Guo, J., Tang, J. *et al.* (2022) Elucidation of the essential oil biosynthetic pathways in *Cinnamomum burmannii* through identification of six terpene synthases. *Plant Science*, **317**, 111203.

Marçais, G. & Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.

Myburg, A.A., Grattapaglia, D., Tuskan, G.A., Hellsten, U., Hayes, R.D., Grimwood, J. *et al.* (2014) The genome of *Eucalyptus grandis*. *Nature*, **510**, 356–362.

National Pharmacopoeia Committee (2020) *Pharmacopoeia of People's Republic of China*. Beijing, China: China Medical Science and Technology Press.

Nawrocki, E.P., Kolbe, D.L. & Eddy, S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.

Nguyen, L.T., Schmidt, H.A., Von, H.A. & Minh, B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, **32**, 268–274.

Ou, S.J., Chen, J.F. & Jiang, N. (2018) Assessing genome assembly quality using the LTR assembly index (LAI). *Nucleic Acids Research*, **46**, e126.

Radwan, A., Kleinwachter, M. & Selmar, D. (2017) Impact of drought stress on specialised metabolism: biosynthesis and the expression of monoterpene synthases in sage (*Salvia officinalis*). *Phytochemistry*, **141**, 20–26.

Ranallo-Benavidez, T.R., Jaron, K.S. & Schatz, M.C. (2020) GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, **11**, 1432.

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C., Vert, J. *et al.* (2015) HiC-pro: an optimized and flexible pipeline for hi-C data processing. *Genome Biology*, **16**, 259.

Slater, G.S. & Birney, E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.

Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.

Tang, C.L., Chen, J., Zhou, Y., Ding, P., He, G., Zhang, L. *et al.* (2021) Exploring antimicrobial mechanism of essential oil of *Amomum villosum* Lour through metabolomics based on gas chromatography-mass spectrometry in methicillin-resistant Staphylococcus aureus. *Microbiological Research*, **242**, 126608.

Tang, L., He, G., Su, J. & Xu, H. (2012) The strategy to promote the development of industry of genuine medicinal material of *Amomum villosum*. *Chinese Agricultural Science Bulletin*, **28**, 94–99.

Testa, A.C., Hane, J.K., Ellwood, S.R. & Oliver, R.P. (2015) CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. *BMC Genomics*, **16**, 170.

Tu, L., Su, P., Zhang, Z., Gao, L., Wang, J., Hu, T. *et al.* (2020) Genome of *Tripterygium wilfordii* and identification of cytochrome P450 involved in triptolide biosynthesis. *Nature Communications*, **11**, 971.

Vranova, E., Coman, D. & Gruissem, W. (2013) Network analysis of the MVA and MEP pathways for isoprenoid synthesis. *Annual Review of Plant Biology*, **64**, 665–700.

Wang, D., Zhang, Y., Zhang, Z., Zhu, J. & Yu, J. (2010) KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics, Proteomics & Bioinformatics*, **8**, 77–80.

Wang, H., Ma, D., Yang, J., Deng, K., Li, M., Ji, X. *et al.* (2018) An integrative volatile terpenoid profiling and transcriptomics analysis for gene mining and functional characterization of AvBPPS and AvPS involved in the monoterpenoid biosynthesis in *Amomum villosum*. *Frontiers in Plant Science*, **9**, 846.

Wang, X., Gao, Y., Wu, X., Wen, X., Li, D., Zhou, H. *et al.* (2021) High-quality evergreen azalea genome reveals tandem duplication-facilitated low-altitude adaptability and floral scent evolution. *Plant Biotechnology Journal*, **19**, 2544–2560.

Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G. *et al.* (2018) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, **35**, 543–548.

Yang, R., Wang, J., Gao, W., Jiang, Y., Su, J., Sun, D. *et al.* (2021) Research on the reproductive biological characteristics of *Amomum villosum Lour*. And *amomum longiligulare* T. L. Wu. *PLoS One*, **16**, e0250335.

Yang, Y., Li, S., Xing, Y., Zhang, Z., Liu, T., Ao, W. *et al.* (2022) The first high-quality chromosomal genome assembly of a medicinal and edible plant *Arctium lappa*. *Molecular Ecology Resources*, **22**, 1493–1507.

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

Yue, J.J., Zhang, S., Zheng, B., Raza, F., Luo, Z., Li, X. *et al.* (2021) Efficacy and mechanism of active fractions in fruit of *Amomum villosum* Lour. For gastric cancer. *Journal of Cancer*, **12**, 5991–5998.

Zhang, H.K., Gao, S.H., Lercher, M.J., Hu, S.N. & Chen, W.H. (2012) EvolView, an online tool for visualizing, annotating and managing phylogenetic trees. *Nucleic Acids Research*, **40**, w569–w572.

Zhao, H., Li, M., Zhao, Y., Lin, X., Liang, H., Wei, J. *et al.* (2021) A comparison of two monoterpenoid synthases reveals molecular mechanisms associated with the difference of bioactive monoterpenoids between *Amomum villosum* and *amomum longiligulare*. *Frontiers in Plant Science*, **12**, 695551.