Supplementary Information for

Multiomics analysis reveals the evolutionary origin of diterpenoid alkaloid biosynthesis pathways in *Aconitum*

Dake Zhao[1,†], Ya Zhang[3,†], Huanxing Ren[2,4], Yana Shi[5], Ding Dong[1], Zonghang Li[1], Guanghong Cui[6], Yong Shen[7], Zongmin Mou[1], Edward J. Kennelly[8,9], Luqi Huang[5,6], Jue Ruan[3], Suiyun Chen[1,*], Diqiu Yu[10,*], Yupeng Cun[2,4,*]

This file includes:

Other Supplementary Material for this manuscript includes the following:

(available at https://cunlab.org/pubs/si4avilmo/)

Supplementary Datasets S1 to S11 (.xlsx)

Materials and methods

1. Plant material, DNA, and RNA preparation

In China, *Aconitum carmichaeli*, *A. vilmorinianum*, and *A. kusnezoffii* are used in traditional Chinese medicine (TCM). Among these species, *A. vilmorinianum* is one of the most important components of the famous medicine "Yunnan Baiyao" (Supplementary Figure S1). The plant is endemic to the central Yunnan Plateau, and *A. vilmorinianum* is a threatened species due to extensive artificial collection. We, therefore, select *A. vilmorinianum* for genome sequencing.

Three individuals from each of the four regions were used to extract RNA from four tissues. Apart from Luquan population, where *A. vilmorinianum* was artificially planted, and the individuals were from field collections. Total RNA was extracted from these tissues using the HiPure Plant RNA Kit (Ambion Trizol). The extracted RNA was reverse transcribed into cDNA using the SMARTer PCR cDNA Synthesis Kit (Clontech), followed by transcriptome sequencing.

2. Genome sequencing, assembly, and map integration of *A. vilmorinianum*

2.1 Estimation of genome size using k-mer analysis

The size of the *A. vilmorinianum* genome was estimated using a k-mer (k = 19) analysis-based approach together with Illumina paired-end short reads. The sequencing depth was estimated by determining the highest peak value in the frequency curve of the K-mer occurrence distribution (Supplementary Figure S2). The final genome size was estimated to be 5.48 Gb based on the K-mer statistics.

2.2 Quality assessment and validation of assembly.

The results of Nanopore data assembly are evaluated from three perspectives. First, the short sequence of Illumina HiSeq was aligned to the reference genome using BWA software. A

total of 2,225 million reads were aligned to the genome, and 1,868 million reads were located to the reference genome, and the distance was consistent with the distribution of sequencing fragments, with an alignment rate of 82.91% (Supplementary Table S7). We employed REDUNDANS (v.013b)[32] to identify heterozygous contigs based on at least 85% identity and overlap of binger than 66% of the shorter sequence in the comparison of each pairwise, and only longer of the two redundant contigs was retained.

Secondly, CEGMA V2.5 was used to assess the integrity of the assembly, and 405 conserved core genes were compared, accounting for 88.43%, of which 171 highly conserved CEGs accounted for 68.95% (Supplementary Table S5). This was finally evaluated using the BUSCO software (BUSCO V5) with the following parameters: --E value 1E-05 (e-value cut off for BLAST searches), --SP Generic. Total of 1828 complete BUSCOs were found, accounting for 86.19%, and 1581 complete and single-copy BUSCOs were found (Supplementary Table S6). A complete *de novo* assembly pipeline was:

a). We employ CANU to correct errors in the raw data of 3rd long Nanopore raw sequencing data and assembled by SMATdenovo;

b). We used Medaka to correct the assembled genome;

c). Pilon was used to applying 2ed generation short sequencing results to polish 3rd generation long sequencing results;

d). Mapping 2ed short and 3rd long sequencing to get the quality of reference genome. and BUSCO was used to evaluate the complete reference genome.

e). Combined with HiC data, scaffold in the reference genome were mapped to chromosome-level genome assembly.

3. Genome annotation

Repetitive elements were estimated to comprise 90.39% of the high-quality assembly genome, including long terminal repeat (LTR) retrotransposons (60.39%), DNA transposable elements (5.02%), and long interspersed nuclear elements (5.28%) (Table S10). Gypsy elements were the most abundant LTRs and comprised 39.90% of the entire genome assembly; Copia repeats comprised 19.50% of the genome assembly.

3.1 Non-coding RNA annotation

According to the structural characteristics of different non-coding RNAs, different strategies were used to predict different non-coding RNAs. Based on the Rfam[1] database, Blastn software was used to identify microRNAs and rRNAs in the whole genome. The tRNAscan-SE [2] software was used for tRNA identification.

3.2 Pseudogene annotation

Using the predicted protein sequences, BLAT[3] was used for alignment to identify homologous gene sequences in the genome, and GeneWise [4] was used to identify immature stop codons and frameshift mutations in gene sequences.

4. Evolutionary genomics analyses

Twenty species' genome assemblies: *Chlamydomonas eustigma, Volvox reticuliferus, Physcomitrella patens, Gnetum montanum, Welwitschia mirabilis, Amborella trichopoda, Oryza sativa, A. vilmorinianum, A. coerulea, Artemisa annua, Macleaya cordata, Papaver somniferum, Arabidopsis thaliana, Jatropha curcas, Hevea brasiliensis, Vitis vinifera, Ricinus communis, Coptis chinensis, Kingdonia uniflora, and Thalictrum thalictroides (Table S17)*

4.1 Detection of gene duplication events

Analysis of systemic blocks between *A. vilmorinianum* and *Aquilegia coerulea* and grape genomes by MCScanX [5], based on all homologous proteins of these three species using an all-to-all search in BLAST [6]. We also found several chromosomal segments duplicated in *A. vilmorinianum*, for example, chromosomes 1 and 3, chromosomes 1 and 7, and chromosomes 2 and 4 (Supplementary Figure S6). In addition, there is an overall 2:1 syntenic relationship (Figure 1-C) between *A. vilmorinianum* and *Aquilegia coerulea* (Figure 1-D). There is a similar 3:4 relationship in the collinearity analysis of grapes and *A. vilmorinianum*, corresponding to one fragment of the grape corresponding to four fragments in *A. vilmorinianum* and one fragment of *A. vilmorinianum* corresponding to three fragments in the grape（Supplementary Figure S5 and Figure S6).

We then used the distribution of synonymous substitution rates per gene (*Ks*) between collinear paralogous genes to identify WGD events. The distribution of *Ks* of collinear paralogous genes from *A. vilmorinianum* and *Aquilegia coerulea* showed that the peak was around 0.82 (Figure 1-D), suggesting the presence of whole genome duplication. Taken together, our analysis provided strong evidence for a single whole genome duplication event in *A. vilmorinianum*.

4.2 Estimate of whole genome duplication timing

To time the WGD of *A. vilmorinianum*, we estimated the average evolutionary rate for Ranunculaceae using *A. vilmorinianum* and *Aquilegia coerulea*. The Ks values were then fitted to a Gaussian mixture model to determine the number of components in the *Ks* distribution [7]. The divergence time of 65.35 million years ago (MYA) between *A. vilmorinianum* and *Aquilegia coerulea* was obtained based on our divergence estimate using MCMCtree (Figure 1-D), we calculated the synonymous substitutions per site per year (r) for Ranunculaceae to be 6.108e-9 (T = Ks / 2r). We dated the WGD (*Ks* = 0.822) to 67.32 ± 5.52 MYA in *A. vilmorinianum*, and calculated the WGD timing to 124.10 ±10.18 MYA in *Aquilegia coerulea*. The estimated timing for the WGD events of *A. vilmorinianum* and *Aquilegia coerulea*, as well as previously reported WGD/WGT in three other angiosperm species (*Macleaya cordata*, *Papaver somniferum*, *Arabidopsis thaliana*) were plotted in the phylogenetic tree (Figure 1-B).

5. Transcriptome analysis

5.1 Gene differential expression analysis

The clean reads were mapped to the *A. vilmorinianum* genome using STAR (v 2.7.1)[8] software, and then the transcripts were quantified using RSEM (v 1.3.2)[9] software. Differential gene analysis was performed using DESeq2 in the four tissues according to two conditions, $|\log2\ \text{FoldChange}| \geq 1$ and $\text{padj} < 0.05$, and 11475 differentially expressed genes were obtained in *A. vilmorinianum*.

5.2 Co-expression network and the minimum spanning tree of candidate genes for diterpenoid alkaloid synthesis

Differentially expressed candidate genes for diterpenoid alkaloid synthesis, a total of 313 genes, including 4 geranylgeranyl pyrophosphate synthases (GGPPSs), 3 *ent*-copalyl diphosphate synthases (CPSs), 4 *ent*-kaurene synthases (KSs), 2 *ent*-kaurene oxidases (KOXs), 36 aminotransferases (ATFs), 142 CYP450s (P450s), 41 *O*-methyltransferases (OMTs), 8 BAHD acyltransferases (BAHDs). The WGCNA package was used to construct a weighted gene co-expression network for the genes of these eight gene families. Then, a minimum spanning tree analysis was then performed on the co-expression network to find hub genes. Several hub genes were found in the minimum spanning tree, consisting of 1 GGPPS (EVM0025196), 1 Glutamate--glyoxylate aminotransferase (EVM0007090), 16 P450s (EVM0038628, EVM0046786, EVM0046628, EVM0023781, EVM0026284, EVM0052907, EVM0002806, EVM0050322, EVM0057232, EVM0050789, EVM0019247, EVM0023399, EVM0055500, EVM0035373, EVM0030489, and EVM0014454), 7 OMTs (EVM0008171, EVM0000759, EVM0023968, EVM0046328, EVM0056152, EVM0021046, EVM0034357, and EVM0005923), 3 BAHDs (EVM0000058, EVM0046469, and EVM0051236).

5.3 Expression of transposable elements and Construct candidate gene and transcription factor

HMMER software was used to make predictions based on HMM models of transcription factor families. 37 types of transcription factors were predicted in *A. vilmorinianum*, and 1748 transcription factors (TFs) were obtained after screening. The top 10 most abundant transcription factor types were: GARP (390), MYB (261), AP2/EREBP(160), bHLH (122), C2C2 (116), NAC (108), ABI3/VP1 (89), HB (81), WRKY (62), and bZIP (66) (Figure 2-D and Supplementary Data S2).

To further understand the genes and regulation associated with diterpenoid alkaloid synthesis in *A. vilmorinianum*, the WGCNA package was used to construct a weighted gene co-expression network for the predicted transcription factors and the genes associated with diterpenoid alkaloid synthesis. Cytoscape software was used to visualize co-expression networks with edge weights greater than 0.24 (Figure 2-D and Supplementary Data S2). We

identified 109 transcription factors that may be involved in the regulation of 83 pathway genes. Several transcription factors such as AP2/ERF, bHLH, and WRKY have a high correlation with pathway genes. However, the regulation of diterpenoid alkaloid biosynthesis in *A. vilmorinianum* may be a relatively complex system, with several unusual TFs also showing strong associations, such as members of the NAC, C3H, and GRAS families. The types and amounts of transcription factors associated with aminotransferase and CYP450 gene families are the largest. In addition, 34 alkaloid synthesis candidate genes were associated with AP2/ERF transcription factors (Supplementary Figure S8 and Figure S9).

6. Metabolome analysis

6.1 Sample preparation and UPLC-MS/MS analysis

A. vilmorinianum plant samples were lyophilized using a lyophilizer (Scientz-100F), then the samples were ground into powders using a mixer mill (MM 400, Retsch) for 1.5 min at 30 Hz. 100 g of powder was dissolved in 1.2 ml of 70% methanol, vortexed for 30 sec every 30 min, vortexed 6 times, then the samples were stored in a refrigerator at 4℃ overnight. The solution was centrifuged at 12000 rpm for 10 minutes, the supernatant was filtered through a 0.22μm pore size membrane to perform UPLC-MS/MS analysis.

The sample extracts were analyzed using a UPLC-ESI-MS/MS system (UPLC, SHIMADZU Nexera X2, www.shimadzu.com.cn; MS, Applied Biosystems 4500 Q TRAP, www.appliedbiosystems.com.cn/). The analytical conditions were as follows, UPLC: column, Agilent SB-C18 (1.8 μm, 2.1 mm*100 mm); the mobile phase consisted of solvent A, pure water with 0.1% formic acid, and solvent B, acetonitrile with 0.1% formic acid. Sample measurements were performed with a gradient program using the starting conditions of 95% A, and 5 % B. Within 9 minutes, a linear gradient to 5% A, 95% B was programmed, and a composition of 5% A, 95% B was maintained for 1 minute. Then, within 1.10 min, a composition of 95% A, and 5.0% B was set and maintained for 2.9 min. The flow rate was set to 0.35 ml per minute; the column oven was set to 40°C; the injection volume was 4 μl. The effluent was alternatively connected to an ESI-triple quadrupole-linear ion trap (QTRAP)-MS.

The operating parameters of the ESI source: one ion source, Turbo Spray; source temperature 550°C; ion spray voltage (IS) 5500 V (positive ion mode)/-4500 V (negative ion mode); ion source gas I (GSI), gas II(GSII), curtain gas (CUR) were set to 50, 60, and 25.0 psi, respectively; the collision-activated dissociation(CAD) was high. Instrument tuning and mass calibration were performed with 10 and 100 μmol/L polypropylene glycol solutions in QQQ and LIT modes, respectively. QQQ scans were acquired as MRM experiments with collision gas (nitrogen) set to medium. DP and CE for individual MRM transitions were performed with further DP and CE optimization. A specific set of MRM transitions was monitored for each period according to the metabolites eluted within this period.

6.2 Data analysis of metabolic profile

Analyst software (v1.6.3) was used to process the mass spectrometry data. Chromatographic peak integration and calibration were performed using MultiaQuant software. Unsupervised principal component analysis (PCA) was performed by the *prcomp* statistics function in R. Hierarchical cluster analysis (HCA) results of samples and metabolites were presented as heatmaps with dendrograms, while Pearson correlation coefficients (PCC) between samples were calculated with the *cor* function in R and presented as heatmaps (Supplementary Figure S17 and Figure S18). Significantly regulated metabolites between groups were determined by VIP $\geqslant$ 1 and absolute Log_2FC (fold change) $\geq$ 1. VIP values were extracted from the OPLS-DA result, which also contains score plots and permutation plots, and were generated using R package MetaboAnalystR. The data was log transform (log_2) and mean centered before OPLS-DA. A permutation test (200 permutations) was performed to avoid overfitting.

6.3 KEGG annotation and enrichment analysis

Identified metabolites were annotated using KEGG Compound database (http://www.kegg.jp/kegg/compound/), and annotated metabolites were then mapped to KEGG Pathway database (http://www.kegg.jp/kegg/pathway.html). Pathways with

significantly regulated metabolites mapped were then fed into MSEA (metabolite sets enrichment analysis); their significance was determined by hypergeometric test's p-values.

6.4 Integrating transcriptome and metabolome data analysis

Based on UPLC-MS/MS technology and metabolite database, alkaloids in four tissues of *A. vilmorinianum* were investigated and 59 diterpene alkaloid metabolites were identified (Supplementary Data S6 and Figure S10). According to the general DA accumulation law, most of the DAs were enriched in the underground part (primary and lateral roots). After normalizing the levels of these compounds accumulated in different tissues and the gene expression of GGPPSs, CPSs, KSs, KOXs, P450s, ATFs, OMTs, and BAHDs gene families, the Pearson correlation coefficient (PCC) was used to calculate the correlation between these compounds and the expression of these genes to construct the gene-metabolite regulatory network. The integrated network demonstrated that the 103 pathway genes contained 12 hub genes (EVM0030489, EVM0052907, EVM0046628, EVM0035373, EVM0046786, EVM0052195, EVM0056152, EVM0021046, EVM0008171, EVM0000058, and EVM0051236) from the minimum spanning tree highly correlated ($p<0.01$) (Supplementary Figure S11 and Data S7). Notably, we also found that two GGPPSs (EVM0025196 and EVM0016299) and one KS (EVM0024293) were highly associated with diterpene alkaloids. In particular, we also found that three CPSs (EVM0013250, EVM0040922, and EVM0029892) and three KOXs genes (EVM0050182, EVM0047597, and EVM0025963) were slightly correlated ($p<0.05$) with diterpenoid alkaloids (Supplementary Figure S11 and Data S7).

7. Skeleton formation of diterpenoid alkaloid biosynthesis

7.1 Identification of diterpenoid alkaloid biosynthesis genes

The putrescine biosynthetic pathway for diterpenoid alkaloid biosynthesis could have three steps: 1) The synthesis of diterpene alkaloids is initiated by the formation of GGPP by three IPPs, followed by the synthesis of diterpenoid skeletons by GGPPSs, CPSs, and KSs; 2)

The skeletons of diterpene alkaloids are synthesized by KOXs and ATFs, which in turn lead to two key steps; 3) Finally, the diterpene alkaloid skeleton is modified by a series of modifying enzymes, such as BAHDs and OMTs, after the oxidation by P450s. According to the above information, several key gene family members involved in the synthesis of diterpene alkaloids were retrieved based on the protein annotation of *A. vilmorinianum* genome in NR, KOG, GO, KEGG, and other functional databases. For TPSs involved in diterpenoid alkaloid synthesis, homologous sequences were searched in *A. vilmorinianum* using BLAST-based on several TPS candidate genes identified in *A. carmichaelii*.

7.2 Phylogenetic analysis of key gene families

Amino acid sequence alignment of five diterpenoid alkaloid biosynthesis gene families (TPSs, P450s, ATFs, BAHD acyltransferases, and OMTs) was performed by MAFFT (1000 bootstrap) [10], and the alignment data were used to construct phylogenetic trees using RAxML[11] with the maximum likelihood method. The phylogenetic trees were visualized using iTOL (https://itol.embl.de/).

We used the terpene synthase genes retrieved in *A. vilmorinianum* genome and various terpene synthases identified in *A. carmichaelii*, *Arabidopsis thaliana*, *Oryza sativa***,** *Salvia miltiorrhiza*, *Tripterygium wilfordii*, and tomato for phylogenetic analysis (Supplementary Data S8). Phylogenetic analysis of terpene synthase (TPS) showed that the TPS family members were mainly distributed in the TPS-A, TPS-B, TPS-C, and TPS-E/F branches in *A. vilmorinianum*. According to the results of phylogenetic tree analysis, we can speculate that the seven genes distributed in TPS-C (pink color) play the function of CPS in the metabolic pathway, catalyzing GGPP to generate ent-CPP. Ten genes distributed in the TPS-E/F branch (blue color) were considered candidate genes for ent-Kaurenes synthesis, catalyzing ent-CPP to produce ent-kaurenes (Supplementary Figure S15).

P450 sequences used to construct phylogenetic trees, including p450 genes annotated in *A. vilmorinianum* genome and a plant p450 database (Plant cytochrome P450 database, PCPD) contains 181 kinds of functions of known plant p450 sequence[12]. In the P450 phylogenetic tree, it can be roughly divided into four branches, the annotated 18 KOX genes forming an

independent small branch with CYP701 (Figure 4-A and Supplementary Figure S13). The aminotransferases identified in *Arabidopsis thaliana* and the genes annotated as aminotransferases in *A. vilmorinianum* were used to construct a phylogenetic tree of aminotransferases (Supplementary Data S9 and Figure S14). The annotated BAHD genes in *A. vilmorinianum* and the BAHD acyltransferases identified in other species were used to construct the phylogenetic tree of BAHD acyltransferases (Figure 4-D and Supplementary Data S10). Plant methyltransferases are divided into C-, N-, S-, and O-methyltransferases. The identified methyltransferases were selected respectively, and the annotated O-methyltransferase gene sequence from *A. vilmorinianum* genome was used to construct a methyltransferase phylogenetic tree (Supplementary Data S11 and Figure S16).

7.3 Evolution of diterpenoid alkaloid biosynthesis genes

In the gene family expansion and contraction analysis, some orthologous genes involved in diterpene alkaloid synthesis were found to be expanded. Surprisingly, the KOX gene family was significantly expanded in *A. vilmorini*anum. Meanwhile, GO enrichment analysis ($p < 0.05$, $FDR < 0.05$) of the expanded and unique genes revealed that these genes were enriched for terpene synthase activity, serine-type peptidase activity, oxidoreductase activity, and O-methyltransferase activity (Supplementary Date S1 and S2), suggesting that the genes involved in diterpene alkaloids may have duplicated and evolved to produce a variety of DAs in *A. vilmorinianum* (Supplementary Figure S19).

Supplementary Tables

Table S1 Survey statistic results of *A. vilmorinianum*.

Table S2 Length distribution of nanopore long reads produced from the *A. vilmorinianum* samples.

Table S3 Sequencing data statistics of *A. vilmorinianum*.

Table S4 Summary of the final genome assembly of *A. vilmorinianum*.

Table S5 Assessment of the gene coverage rate using CEGMA.

Table S6 Assessment of the gene coverage rate using BUSCO.

Table S7 The integrity of the assembled genome was assessed using Illumina sequencing data.

Table S8 Statistics of *A. vilmorinianum* genome after assembly by Hi-C data.

Table S9 Genome statistics after Hi-C assembly

Table S10 Summary of Repeat elements in *A. vilmorinianum* genome.

Table S11 Summary of predicted protein-coding genes of *A. vilmorinianum* genome.

Table S12 Summary of functional annotation of predicted genes of *A. vilmorinianum* genome.

Table S13 Basic statistical results of predicted protein-coding genes of *A. vilmorinianum* genome.

Table S14 The statistical results of non-coding RNA of *A. vilmorinianum* genome.

Table S15 Summary of RNA sequencing data of 24 samples from 3 regions of *A. vilmorinianum*.

Table S16 The statistical results of pseudogene annotation of *A. vilmorinianum* genome.

Table S17 Genomes of 12 plant species used for comparative genomic analysis.

Table S18 The distribution of *A. vilmorinianum* genes among the syntenic regions.

Table S19 Comparison of the gene families among *A. vilmorinianum* and other species.

Table S1. Survey statistic results of *A. vilmorinianum*.

K-mer	K-mer number	K-mer depth	Genome size (Gb)	Revised genome size (Gb)	Heterozygous Ratio (%)	Repeat ratio (%)
19	277,326,647,402	50	5.48	5.76	0.31	80.52

Table S2. Length distribution of nanopore long reads produced from the *A. vilmorinianum* samples.

Length	Reads Number	Total Length	Percentage	Average Length
2000-5000	1,485,735	5,104,567,874	0.86%	3,435.71
5000-10000	2,146,285	16,008,051,522	2.71%	7,458.49
10000-20000	3,883,206	58,486,153,040	9.91%	15,061.30
20000-30000	6,122,731	155,054,822,924	26.28%	25,324.45
30000-40000	4,652,326	160,130,654,962	27.14%	34,419.48
40000-50000	2,214,890	98,278,308,199	16.66%	44,371.64
50000-60000	971,126	52,680,392,820	8.93%	54,246.71
60000-70000	376,393	24,167,957,348	4.09%	64,209.36
70000-80000	145,622	10,809,194,549	1.83%	74,227.75
>=80000	98,891	9,104,024,442	1.54%	92,061.20

Table S3. Sequencing data statistics of *A. vilmorinianum*.

Read type	Insert size	Total data (Gb)	Reads length (bp)	Sequence coverage (X)
Illumina pair-end reads	350 bp	330.72	150	57.42
Nanopore subreads	33.31Kb	589.82	-	102.42
Hi-Cpair-end reads	300 bp	324.19	150	56.28

Table S4. Summary of the final genome assembly of *A. vilmorinianum*.

Contig number	Contig length (bp)	Contig N50 (bp)	Contig N90 (bp)	GC content (%)
26,965	5,758,762,916	311,288	91,698	39.24

Table S5. Assessment of the gene coverage rate using CEGMA.

Number of 458 CEG*	% of 458 CEGs	Number of 248 highly conserved CEGs	% of 248 highly conserved CEGs
405	88.43	171	68.95

Table S6. Assessment of the gene coverage rate using BUSCO.

Complete BUSCOs	Complete and single-copy BUSCOs	Complete and duplicated BUSCOs	Fragmented BUSCOs	Missing BUSCOs	Total Lineage BUSCOs
2,144	1,733	411	36	146	2,326

Table S7. The integrity of the assembled genome was assessed using Illumina sequencing data.

Total reads	Mapped reads	Mapped (%)	Properly mapped reads	Properly mapped (%)
2253863785	2225834300	98.76	1868608322	82.91

Note: Properly mapped (%) means that the paired-end sequencing sequences are all mapped to the reference genome and the distance is consistent with the length distribution of the sequenced fragments.

Table S8. Statistics of *A. vilmorinianum* genome after assembly by Hi-C data.

Group	Cluster Num	Cluster Len (bp)	Order Num	Order Len (bp)
LG01	6098	1297465597	4632	1144532470
LG02	5634	1129901442	4430	1015800089
LG03	2461	586563584	1961	530433242
LG04	2466	512929125	1798	435539651
LG05	2439	546252636	1894	492221476
LG06	2265	482350638	1755	427219531
LG07	1923	591290056	1501	538186255
LG08	1677	355638921	1138	301771692
Total (Ratio %)	24963 (92.34)	5502391999 (95.55)	19109 (76.55)	4885704406 (88.79)

Table S9. Genome statistics after Hi-C assembly

Scaffold Number	Scaffold N50 (bp)	Scaffold N90 (bp)	Contig number	Contig N50 (bp)	Contig N90 (bp)	GC content (%)
7,934	**530,629,242**	**181,968**	**27,035**	**311,288**	**91,698**	**39.24**

Table S10. Summary of Repeat elements in *A. vilmorinianum* genome.

Type	Length (bp)	Percent (%)
Simple Sequence Repeat	882,857	0.02
ClassI	4,818,240,662	83.64
DIRS	316,149,476	5.49
LARD	1,376,213,554	23.89
LINE	303,918,642	5.28
PLE	24,411,826	0.42
TRIM	10,291,147	0.18
SINE	567	0.00
LTR/Copia	1,123,565,291	19.50
LTR/Gypsy	2,298,478,150	39.90
LTR /Unknown	56,960,187	0.99
Unknown	388,617	0.01
ClassII	2,890,73,856	5.02
PotentialHostGene	12,111,436	0.21
Unknown	428,118,014	7.43
Total	5,207,179,061	90.39

Table S11. Summary of predicted protein-coding genes of *A. vilmorinianum* genome.

Method		Gene number
Ab initio	Augustus	56061
	GlimmerHMM	184098
	GeneID	162055
	SNAP	107761
	Genscan	73867
Homolog	*Arabidopsis thaliana*	32680
	Macleaya cordata	37722
	Aquilegia coerulea	37424
	Chinese Lotus	36786
RNAseq	TransDecoder	139790
	GeneMarkS-T	71126
	PASA	41544
Integration	EVM	60351

Table S12. Summary of functional annotation of predicted genes of *A. vilmorinianum* genome.

Database	Annotated number	Percentage(%)
COG	27111	44.92
GO	23863	39.54
KEGG	16293	27.00
KOG	26824	44.45
Pfam	44221	73.27
Swissprot	32588	54.00
TrEMB	46840	77.61
NR	48002	79.54
NT	41234	68.32
Total	52611	87.18

Table S13. Basic statistical results of predicted protein-coding genes of *A. vilmorinianum* genome.

Gene Num	Gene Length	Average Gene length (bp)	Exon Number	Average exon length (bp)	Intron Number	Average intron length (bp)
60351	354731451	5877.81	240351	303.87	240350	1172.03

Table S14. The statistical results of non-coding RNA of *A. vilmorinianum* genome.

RNA classification	Number	Family
miRNA	135	19
rRNA	1045	4
tRNA	1576	25

Table S15. Summary of RNA sequencing data of 24 samples from 3 regions of *A. vilmorinianum*.

Sample	Reads number	Q20(%)	Q30(%)	GC(%)
GJ_1_Y	36,741,142	97.35	92.93	47.57
GJ_1_J	28,204,173	97.64	93.49	45.98
GJ_1_Z	30,339,819	97.25	92.58	45.67
GJ_1_C	33,636,292	97.91	94	45.51
GJ_2_Y	29,354,366	97.54	93.2	45.15
GJ_2_J	32,567,134	97.65	93.49	45.1
GJ_2_Z	29,650,688	97.29	92.82	45.1
GJ_2_C	26,519,157	97.76	93.7	45.61
GJ_3_Y	30,089,707	97.41	92.95	45.38
GJ_3_J	26,834,444	97.18	92.75	46.78
GJ_3_Z	31,994,460	97.47	93.24	45.47
GJ_3_C	31,558,880	97.4	93.07	45.59
LQ_2_Y	27,006,274	97.48	93.16	45.77
LQ_2_J	28,865,777	97.36	93	46.17
LQ_2_Z	33,604,163	97.32	92.91	45.37
LQ_2_C	38,675,220	97.67	93.42	45.55
QJ_1_Y	35,414,845	97.51	93.2	46.12
QJ_1_J	32,651,022	97.33	92.94	45.29
QJ_1_Z	34,148,937	97.47	93.15	45.65
QJ_1_C	28,473,825	97.43	93.11	45.99
QJ_3_Y	33,423,530	97.66	93.53	46.03
QJ_3_J	29,778,213	97.46	93.14	45.17
QJ_3_Z	33,412,717	97.38	92.94	45.15
QJ_3_C	36,187,575	97.48	93.05	46.31

Table S16. The statistical results of pseudogene annotation of *A. vilmorinianum* genome.

Number	Total_len	Average_len
8444	24044163	2847.48

Table S17. Genomes of 21 plant species used for comparative genomic analysis.

Species	NCBI/Phytozome	Reference
A. vilmorinianum	PRJCA014109	This study
Aquilegia coerulea	GCA002738505.1	https://www.uniprot.org/proteomes/UP000230069
Artemisa annua	GCA003112345.1	Shen et al. 2018 [13]
Macleaya cordata	GCA002174775.1	Liu et al. 2017 [14]
Papaver somniferum	GCA003573695.1	Guo et al. 2018 [15]
Arabidopsis thaliana	https://genome.jgi.doe.gov/arabidopsis	Lamesch et al. 2014 [16]
Jatropha curcas	GCA000696525.1	Zhang et al. 2014 [17]
Hevea brasiliensis	GCF001654055.1	Tang et al. 2016 [18]
Vitis vinifera	https://genome.jgi.doe.gov/grape	Jaillon et al. 2007 [19]
Ricinus communis	https://genome.jgi.doe.gov/Rcommunis	Chan et al. 2010 [20]
Oryza sativa	https://genome.jgi.doe.gov/rice	Sun et al. 2007 [21]
Amborella trichopoda	https://genome.jgi.doe.gov/Atrichopoda	Albert et al. 2014 [22]
Coptis chinensis	GCA_015680905.1	Liu et al. 2021 [23]
Gnetum montanum	GCA_015680685.1	Wan et al. 2018 [24]
Kingdonia uniflora	GCA_014058105.1	Sun et al. 2020 [25]
Physcomitrium patens	GCA_000002425.2	Lang et al. 2018 [26]
Thalictrum thalictroides	GCA_013358455.1	Arias et al. 2021 [27]
Welwitschia mirabilis	https://datadryad.org/stash/dataset/doi:10.5061/dryad.ht76hdrdr	Wan et al. 2021 [28]
Chlamydomonas eustigma	GCA_002335675.1	Hirooka et al. 2017 [29]
Volvox reticuliferus	GCA_019650235.1	Yamamoto et al.2021[30]
Selaginella moellendorffii	GCA_000143415.2	Banks et al. 2011[31]

Table S18. The distribution of *A. vilmorinianum* genes among the syntenic regions.

Types of gene duplication	**Genes numbers**	**Percentage**
Singleton	313	9.42
Dispersed	453	13.62
Proximal	127	3.82
Tandem	88	2.65
WGD/segmental	2344	70.5

Note: Singleton: no duplication; Tandem: consecutive duplication; Proximal: duplications in nearby chromosomal region but not adjacent; Dispersed: duplications of modes other than tandem, proximal or WGD/segmental.

Table S19. Comparison of the gene families among *A. vilmorinianum* and other species.

Plant species	Genes number	Genes number in families	Unclustered genes	Family number	Unique Gene number	Unique families number
A. vilmorinianum	60351	52917	7434	15682	15267	3419
Macleaya cordata	21911	21231	680	10720	416	117
Papaver somniferum	84177	61160	1773	12662	13873	1771
Artemisa annua	63226	59777	3449	14093	16929	2500
Arabidopsis thaliana	27416	25585	1831	10637	2714	549
Vitis vinifera	26346	23130	3216	10891	1405	437
Ricinus communis	31221	25091	6130	12828	1729	552
Jatropha curcas	21804	21668	136	10811	310	71
Hevea brasiliensis	35267	34792	475	11329	983	195
Oryza sativa	42189	35137	5494	12039	11263	2038
Amborella trichopoda	26846	22925	3921	11341	4163	823
Coptis chinensis	40011	38509	1502	12938	4856	1062
Gnetum montanum	27491	25295	2059	10782	3347	716
Kingdonia uniflora	43300	41373	1927	12372	9015	1195
Physcomitrium patens	48022	46653	1369	10026	13349	2270
Thalictrum thalictroides	33624	30600	3024	13024	2704	750
Aquilegia coerulea	30023	27178	2845	12166	1313	433
Welwitschia mirabilis	26990	24917	2073	11261	3342	1219
Chlamydomonas eustigma	14161	11947	2214	8015	2268	637
Volvox reticuliferus	22374	19270	3104	8898	6769	1885

Supporting information

Figure S1 A plant of *A. vilmorinianum*.

Figure S2 K-mer analysis for estimating the genome size of *A. vilmorinianum*.

Figure S3 The Hi-C map of *A. vilmorinianum* genome assembly.

Figure S4 The distribution of genes in different species.

Figure S5 Synteny analysis of *A. vilmorinianum*, *A. coerulea,* and *V. vinifera*.

Figure S6 A Dot plot of the synteny block of the *A. vilmorinianum* genomes.

Figure S7 Dot plot of the comparative analysis of the *A. vilmorinianum* and *V. vinifera* by MCScanX.

Figure S8 Heatmap of transcription factors (TF) highly associated with pathway genes in the diterpenoid alkaloid regulatory network in *A. vilmorinianum*.

Figure S9 Heatmap of AP2/ERF transcription factors highly associated with pathway genes in the diterpenoid alkaloid regulatory network in *A. vilmorinianum*.

Figure S10 Chemical structural formulas of diterpenoid alkaloids.

Figure S11 Minimum spanning tree of the integration of gene-metabolite correlation networks with minimum spanning trees in Fig 3-A.

Figure S12 A: A correlation network analysis of pathway genes and diterpenoid alkaloids in *A. vilmorinianum* ($p<0.01$).B: A correlation network analysis of KOXs and diterpenoid alkaloids in *A. vilmorinianum* ($p<0.05$).

Figure S13 Maximum likelihood phylogenetic tree of Cytochrome P450 genes.

Figure S14 Maximum likelihood phylogenetic tree of ATF genes.

Figure S15 Phylogenetic tree of TPS involved in diterpene alkaloid synthesis.

Figure S16 Phylogenetic tree of O-methyltransferase involved in diterpene alkaloid synthesis.

Figure S17 Principal component analysis of metabolic profiles in four tissues of *A. vilmorinianum*.

Figure S18 Heatmap of all metabolic compounds in different tissues of *A. vilmorinianum*.

Figure S19 Expansion and contraction of orthologrous involving diterpene alkaloids in *A. vilmorinianum*.

Figure S1. A plant of *A. vilmorinianum*. A) The young plant. B) The leaves and stem of the adult plant. C) Primary and lateral roots in adult plants.

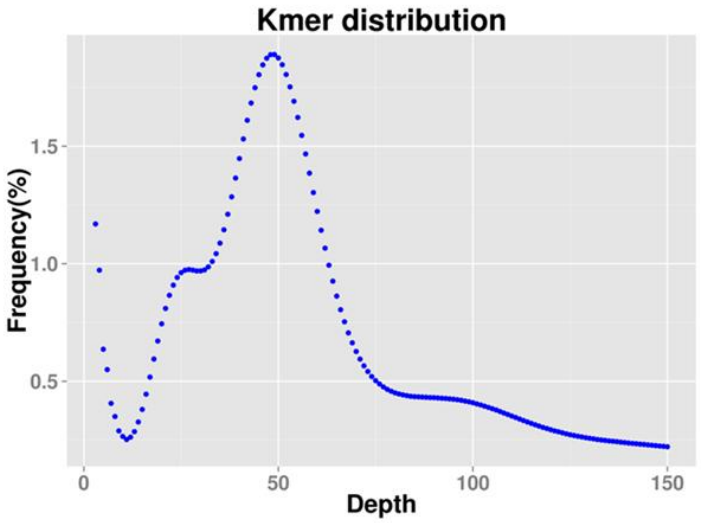

Figure S2. K-mer analysis for estimating the genome size of *A. vilmorinianum*

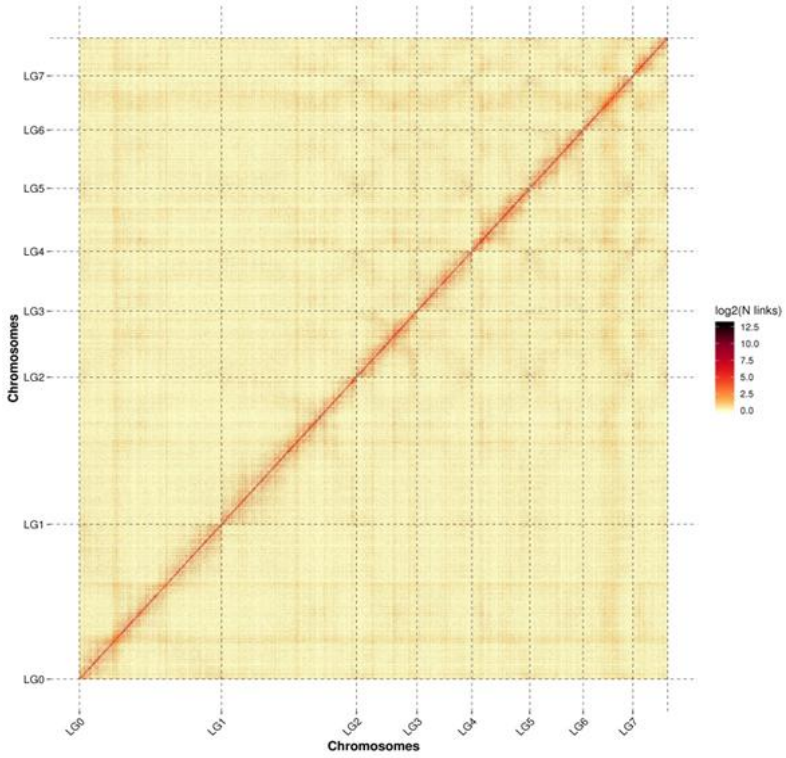

Figure S3. The Hi-C map of *A. vilmorinianum* genome assembly. Hi-C map of the *A. vilmorinianum* genome showing genome-wide all by all interactions. The map shows a high resolution of individual chromosomes that are scaffolded and assembled independently.

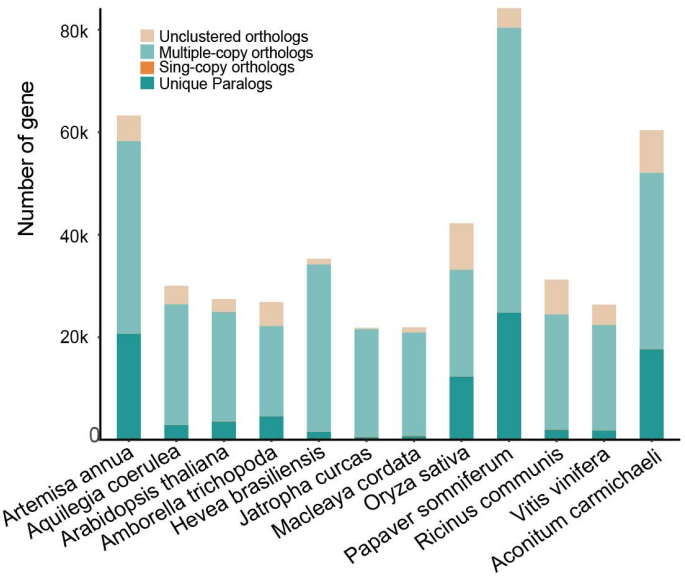

Figure S4. The distribution of genes in different species.

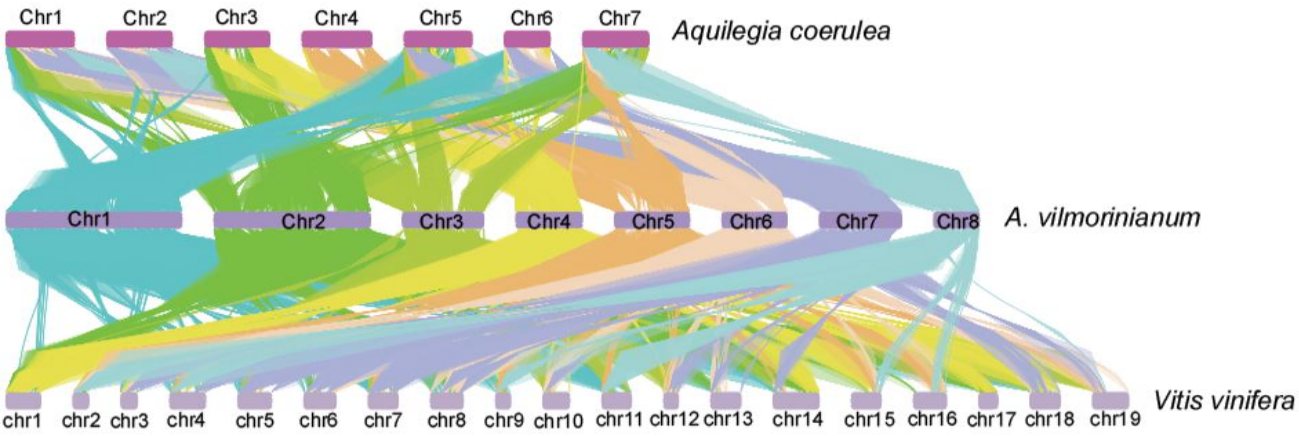

Figure S5. Synteny analysis of *A. vilmorinianum*, *A. coerulea,* and *V. vinifera*.

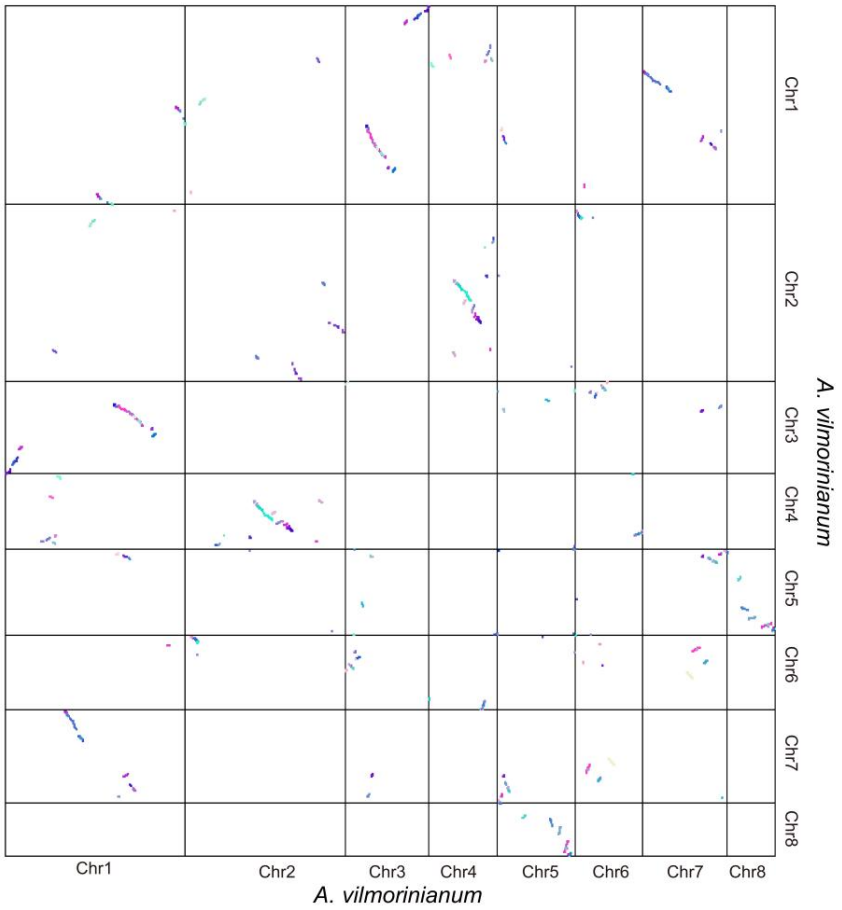

Figure S6. A Dot plot of the synteny block of the *A. vilmorinianum* genomes.

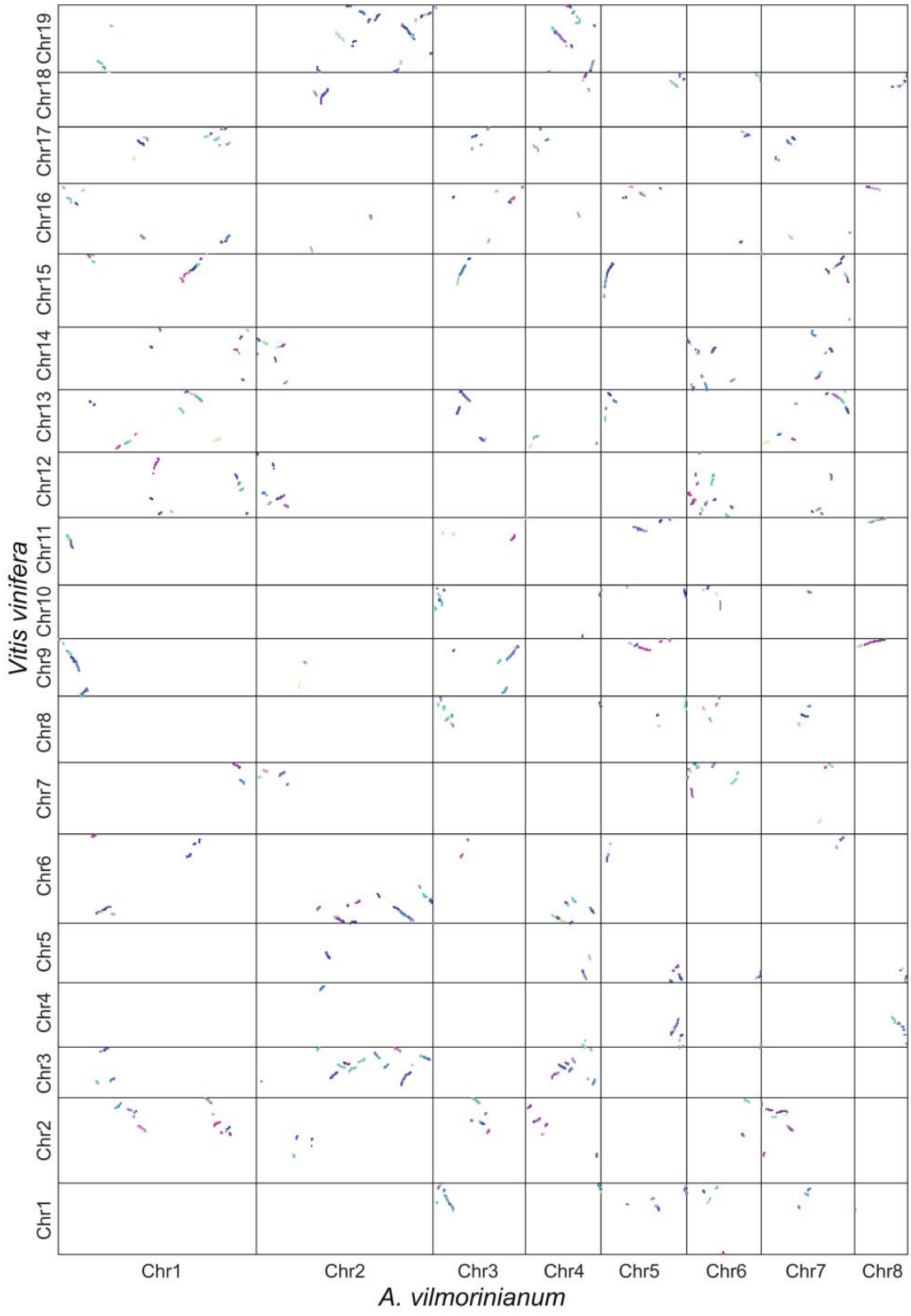

Figure S7. Dot plot of the comparative analysis of the *A. vilmorinianum* and *V. vinifera* by MCScanX.

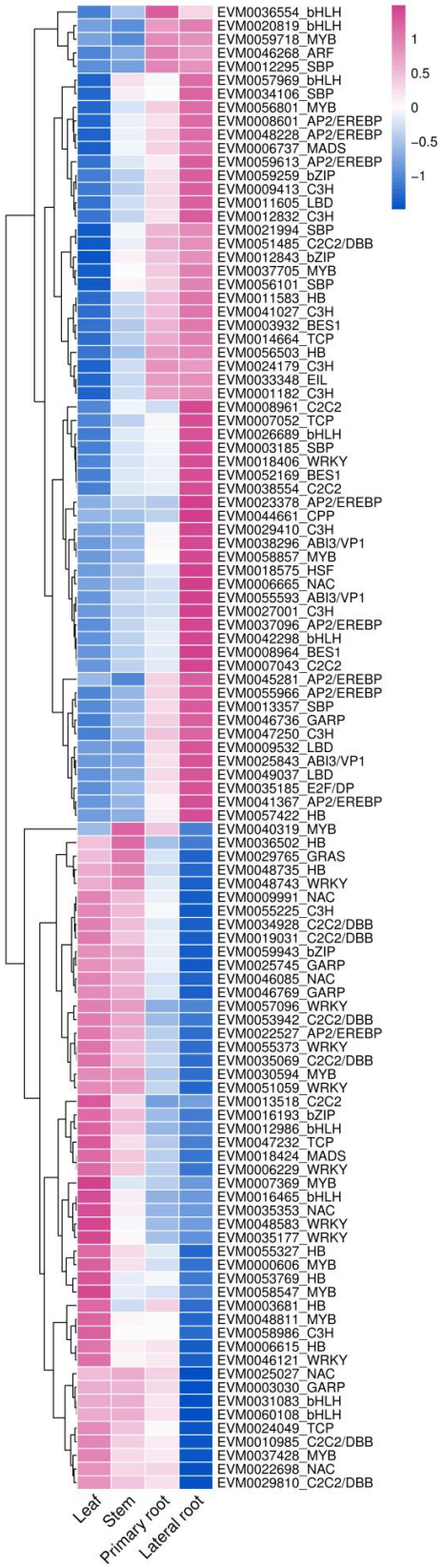

Figure S8. Heatmap of transcription factors (TF) highly associated with pathway genes in the diterpenoid alkaloid regulatory network in *A. vilmorinianum*.

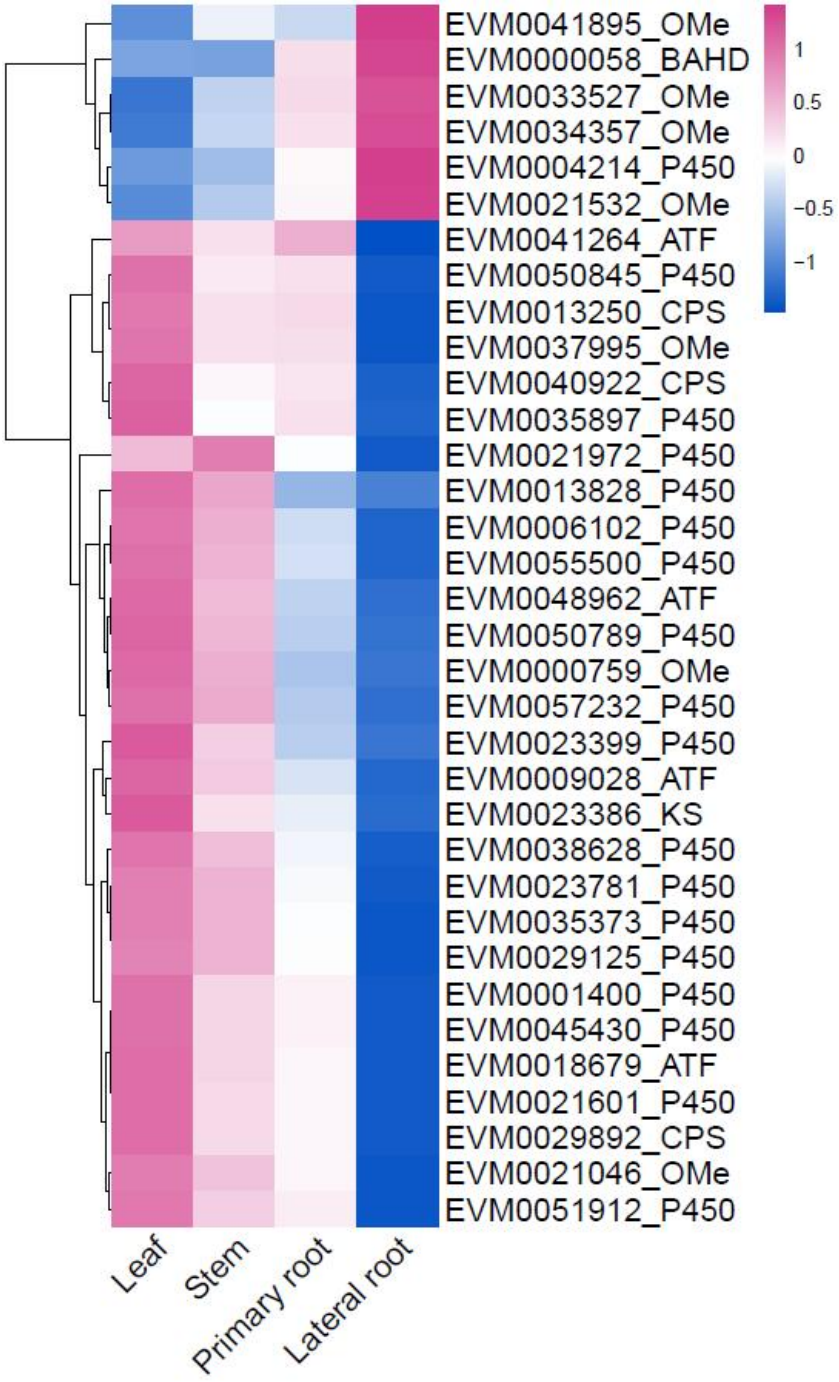

Figure S9. Heatmap of AP2/ERF transcription factors highly associated with pathway genes in the diterpenoid alkaloid regulatory network in *A. vilmorinianum*.

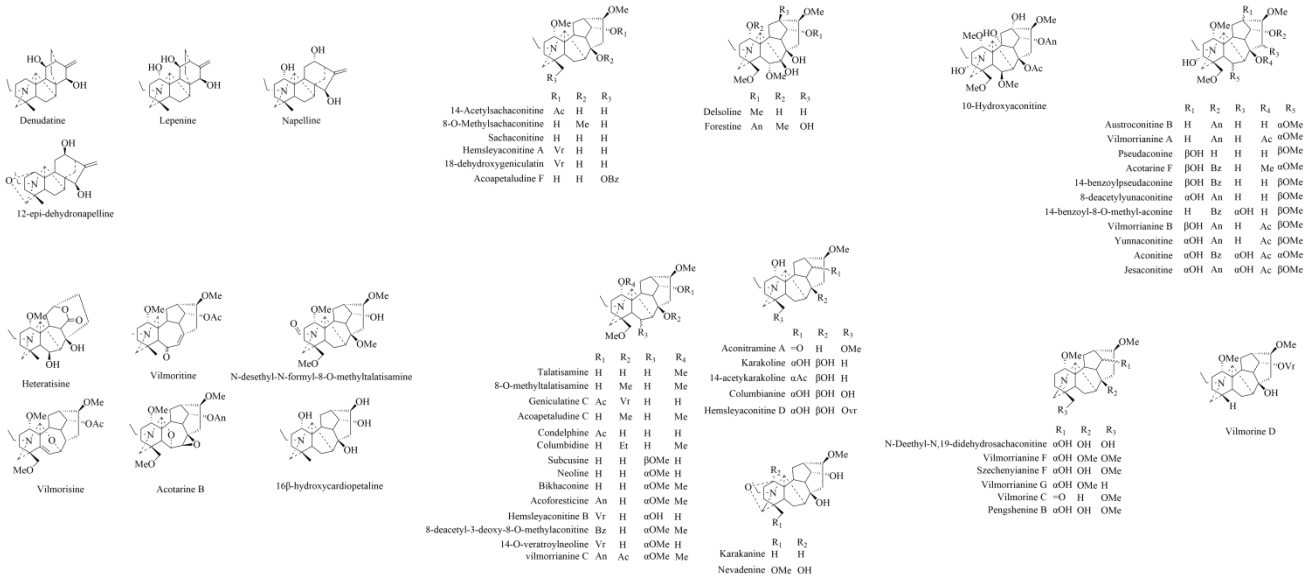

Figure S10. Chemical structural formulas of diterpenoid alkaloids.

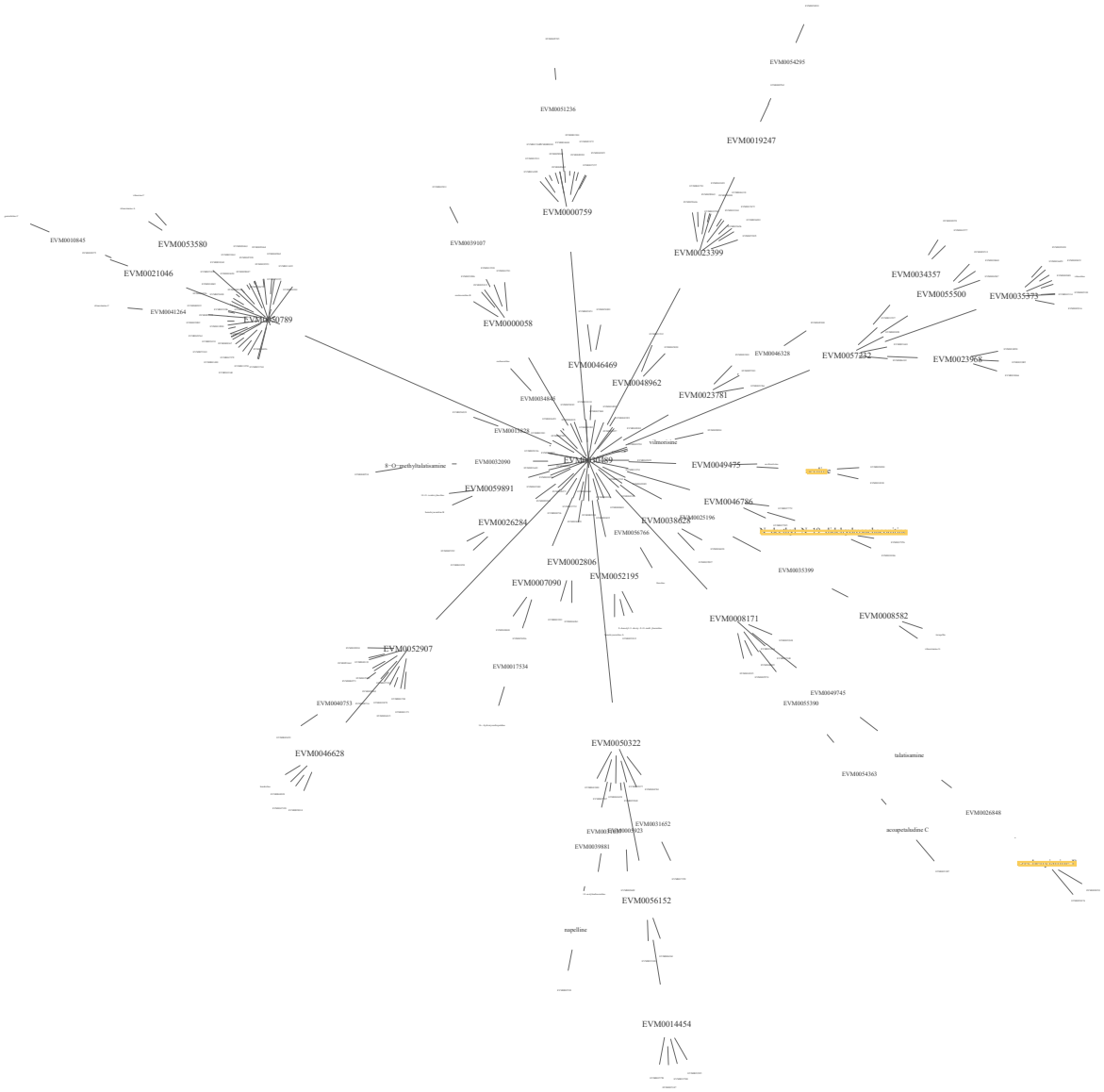

Figure S11. Minimum spanning tree of the integration of gene-metabolite correlation networks with minimum spanning trees in Fig 3-A.

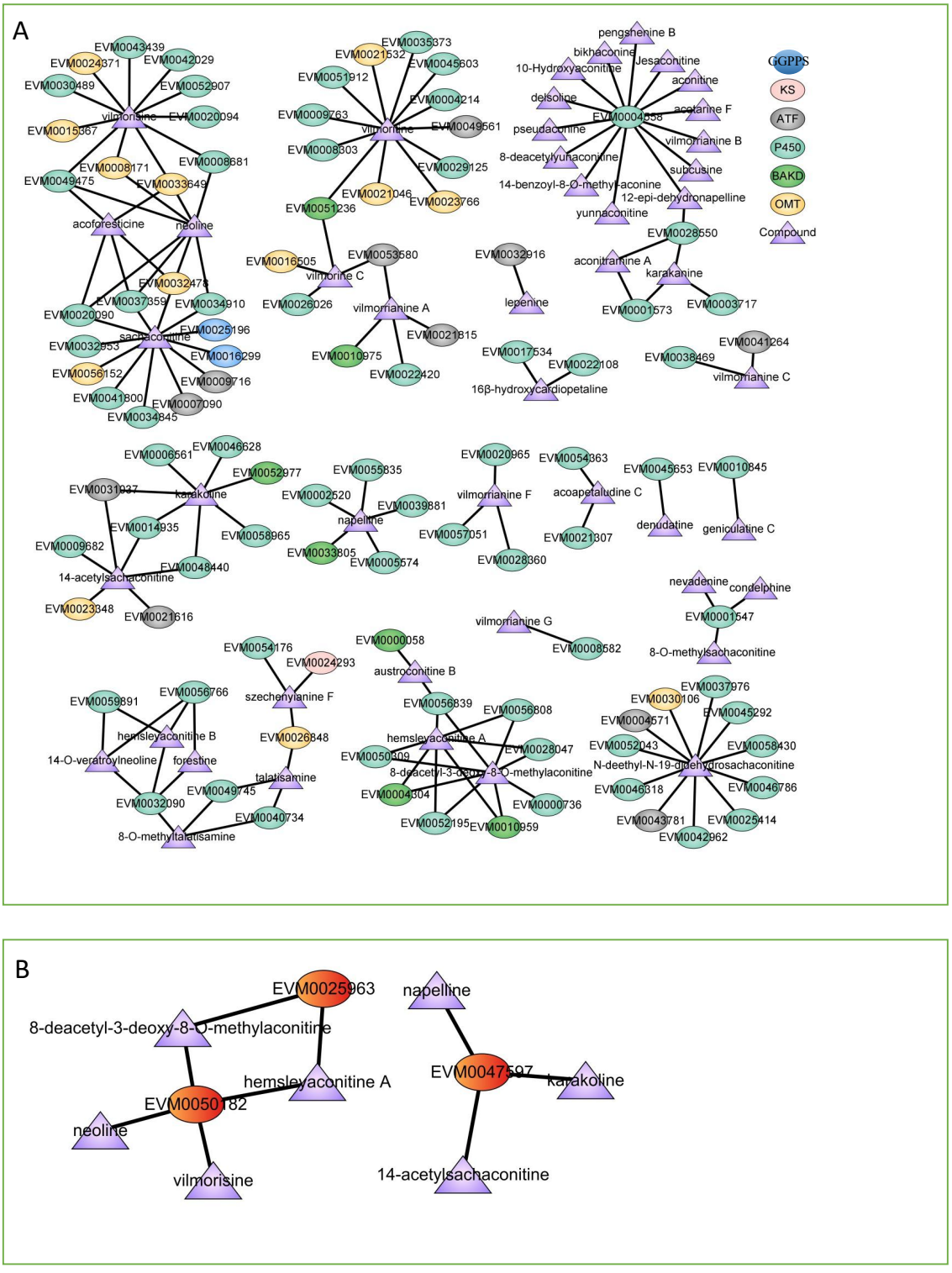

Figure S12. A: Correlation network analysis of pathway genes and diterpenoid alkaloids in *A. vilmorinianum* ($p<0.01$). B: Correlation network analysis of KOXs and diterpenoid alkaloids in *A. vilmorinianum* ($p<0.05$).

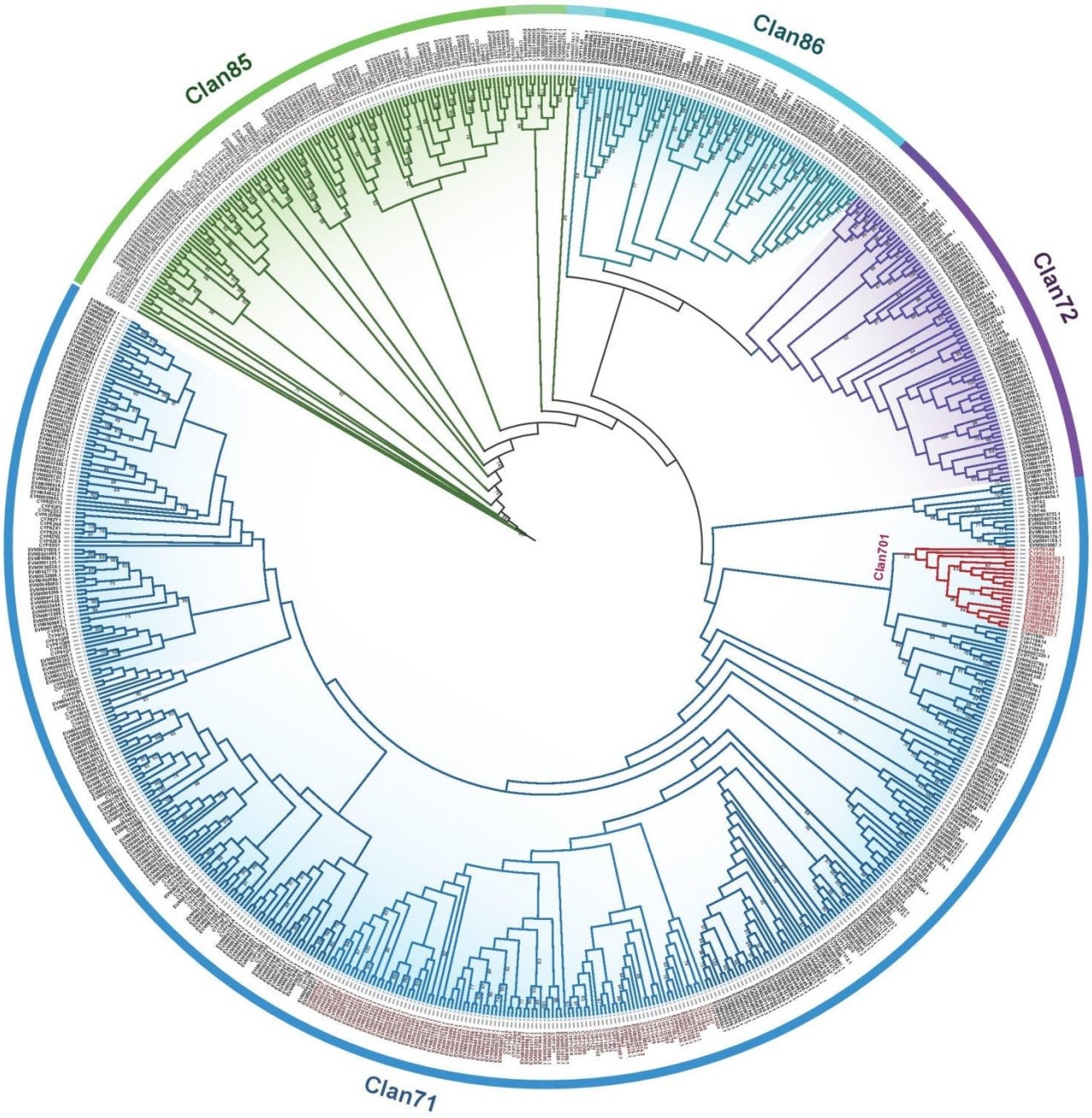

Figure S13. Maximum likelihood phylogenetic tree of Cytochrome P450 genes. The ML tree was constructed using RaxML(v 8.2.12) with the PROTGAMMAJTT and 1000 bootstrap replicates. In summary, the phylogenetic tree of Cytochrome P450 can be roughly divided into four branches, including Clan85, Clan86, Clan71, and Clan72.

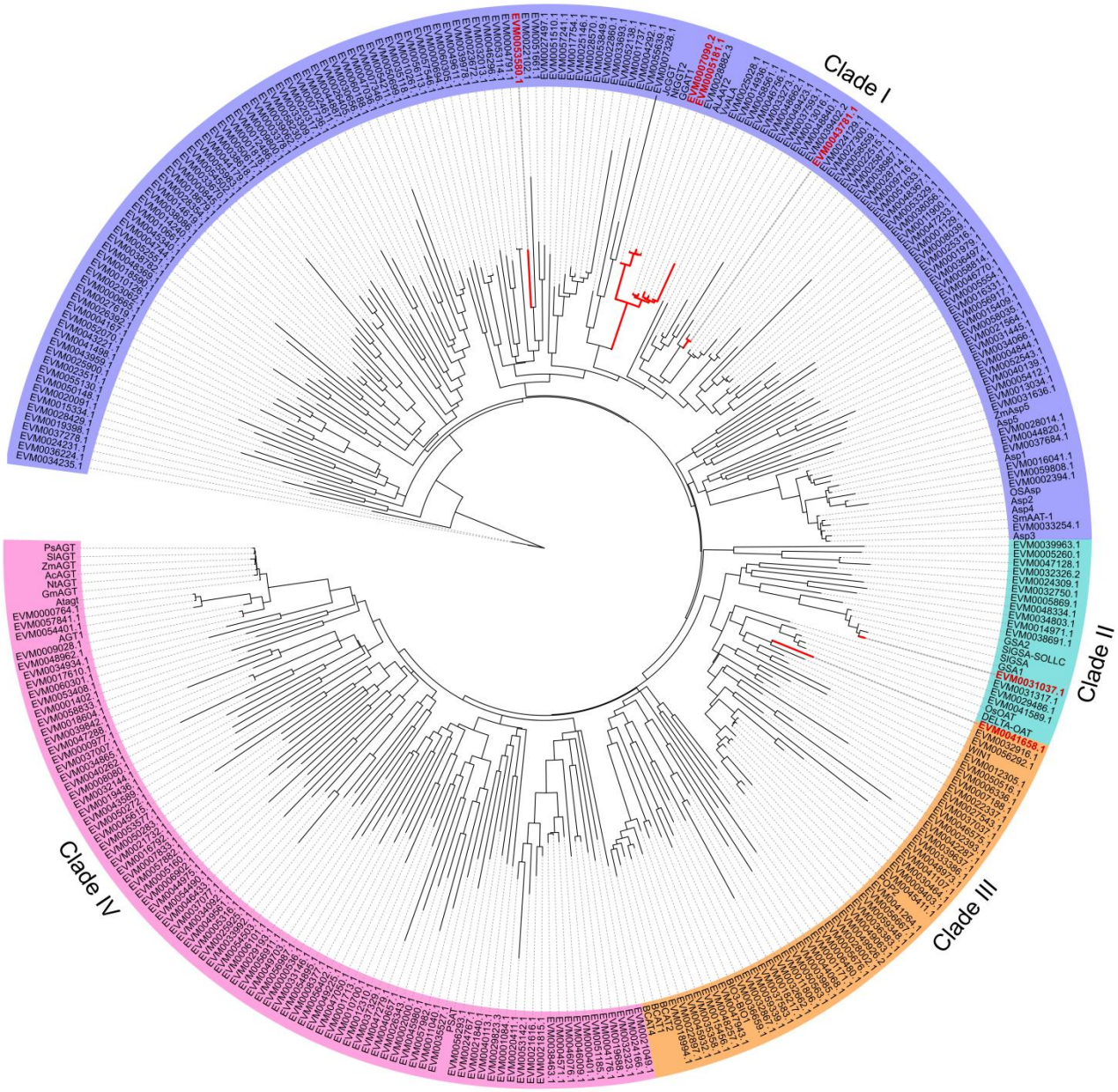

Figure S14. Maximum likelihood phylogenetic tree of ATF genes. The ML tree was constructed using RaxML(v 8.2.12) with the PROTGAMMAJTT and 1000 bootstrap replicates. In summary, the phylogenetic tree of ATF can be roughly divided into four branches, including Clade I, Clade II, Clade III, and Clade IV.

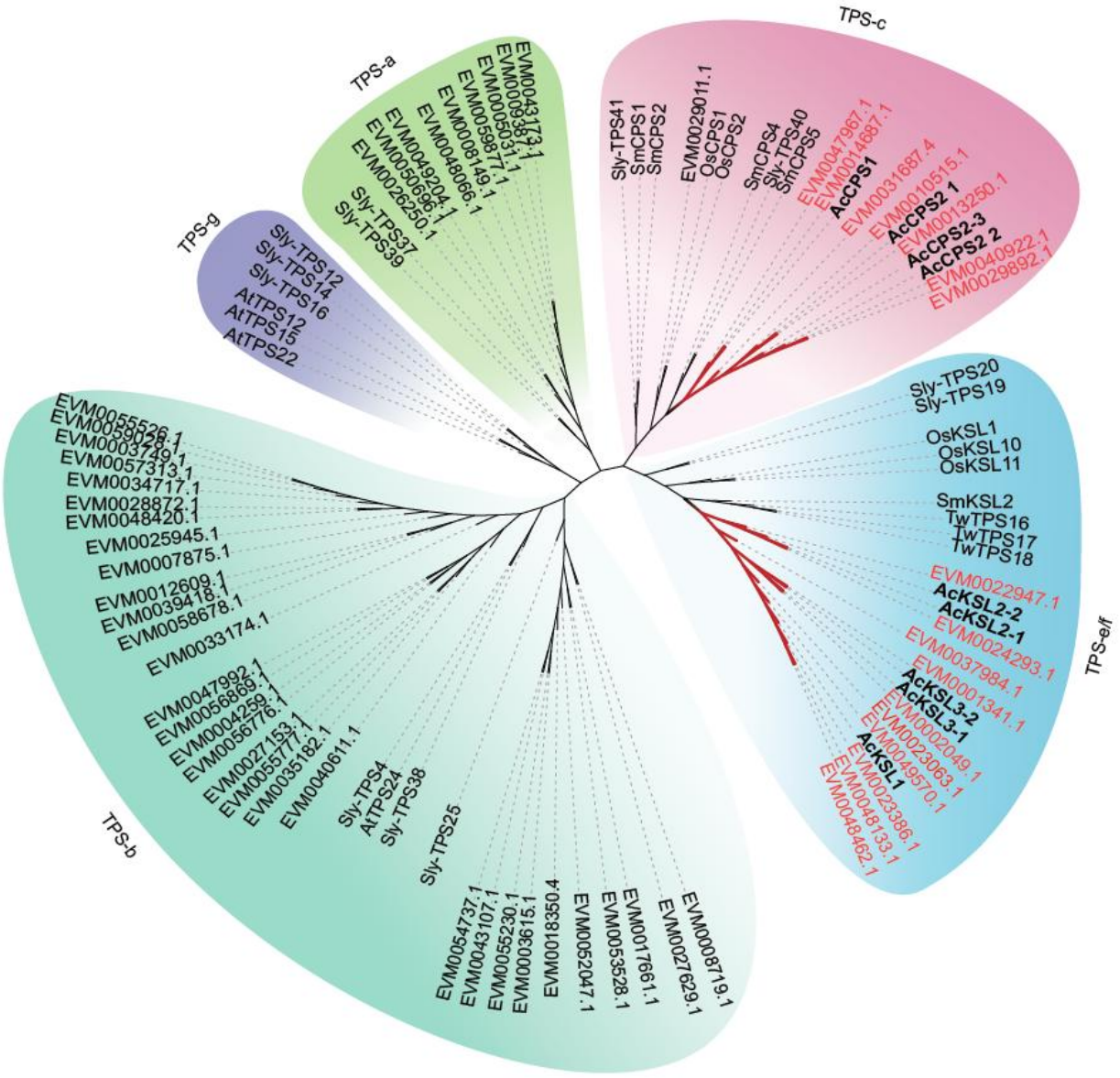

Figure S15. Phylogenetic tree of TPS involved in diterpene alkaloid synthesis. The maximum likelihood tree was reconstructed using RAxML. Numbers on branches indicate the bootstrap percentage values calculated from 1000 bootstrap replicates. TPSs from representative characterized subfamily genes are shown in Data S6. Copalyl diphosphate synthase/ ent-kaurene synthase in Arabidopsis and tomato were used as outgroups.

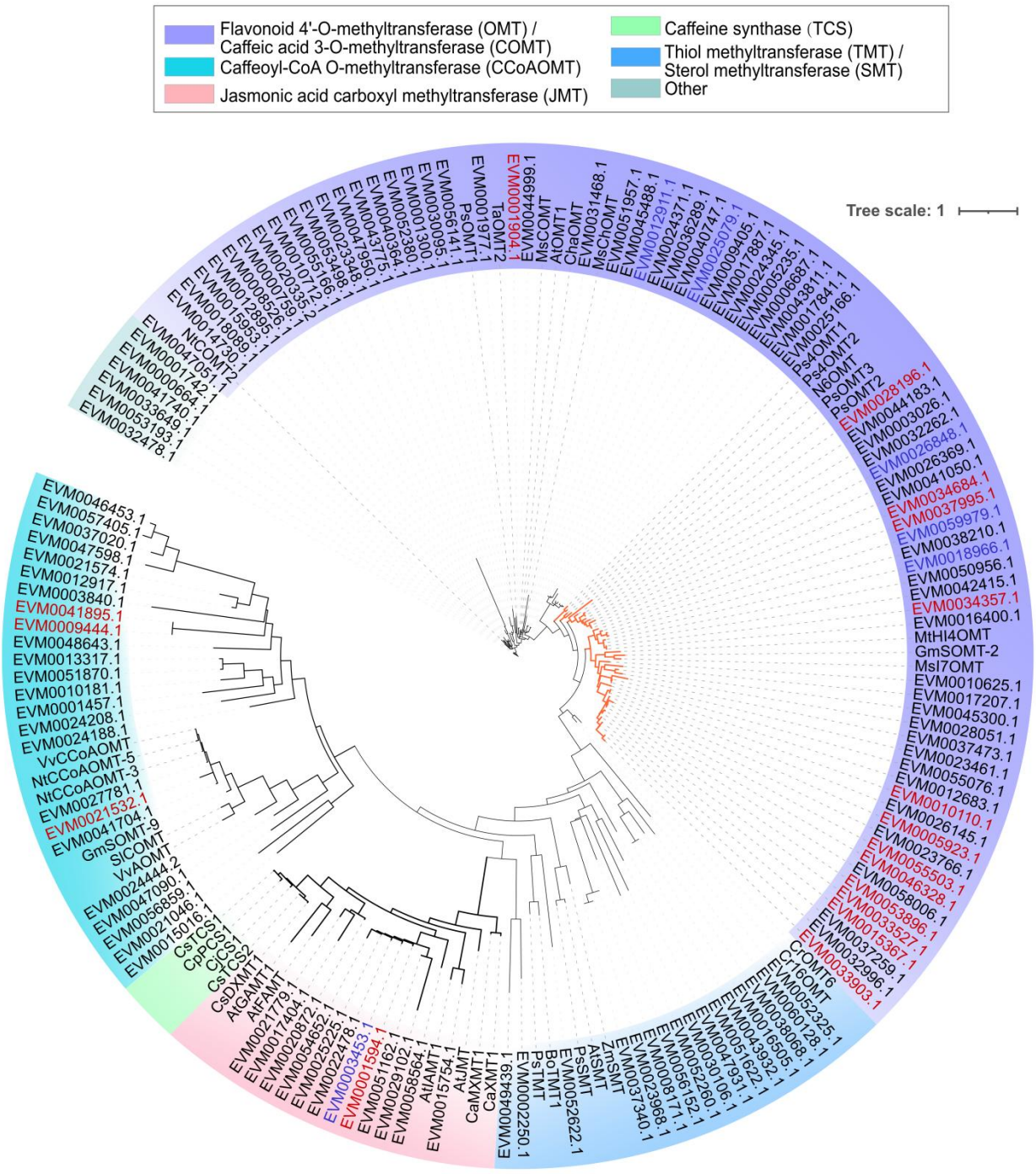

Figure S16. Phylogenetic tree of O-methyltransferase involved in diterpene alkaloid synthesis. The maximum likelihood tree was reconstructed using RAxML(1000 bootstrap). The 19 OMT genes (red) identified in the gene-metabolite correlation analysis were distributed in different branches of the OMT phylogenetic tree.

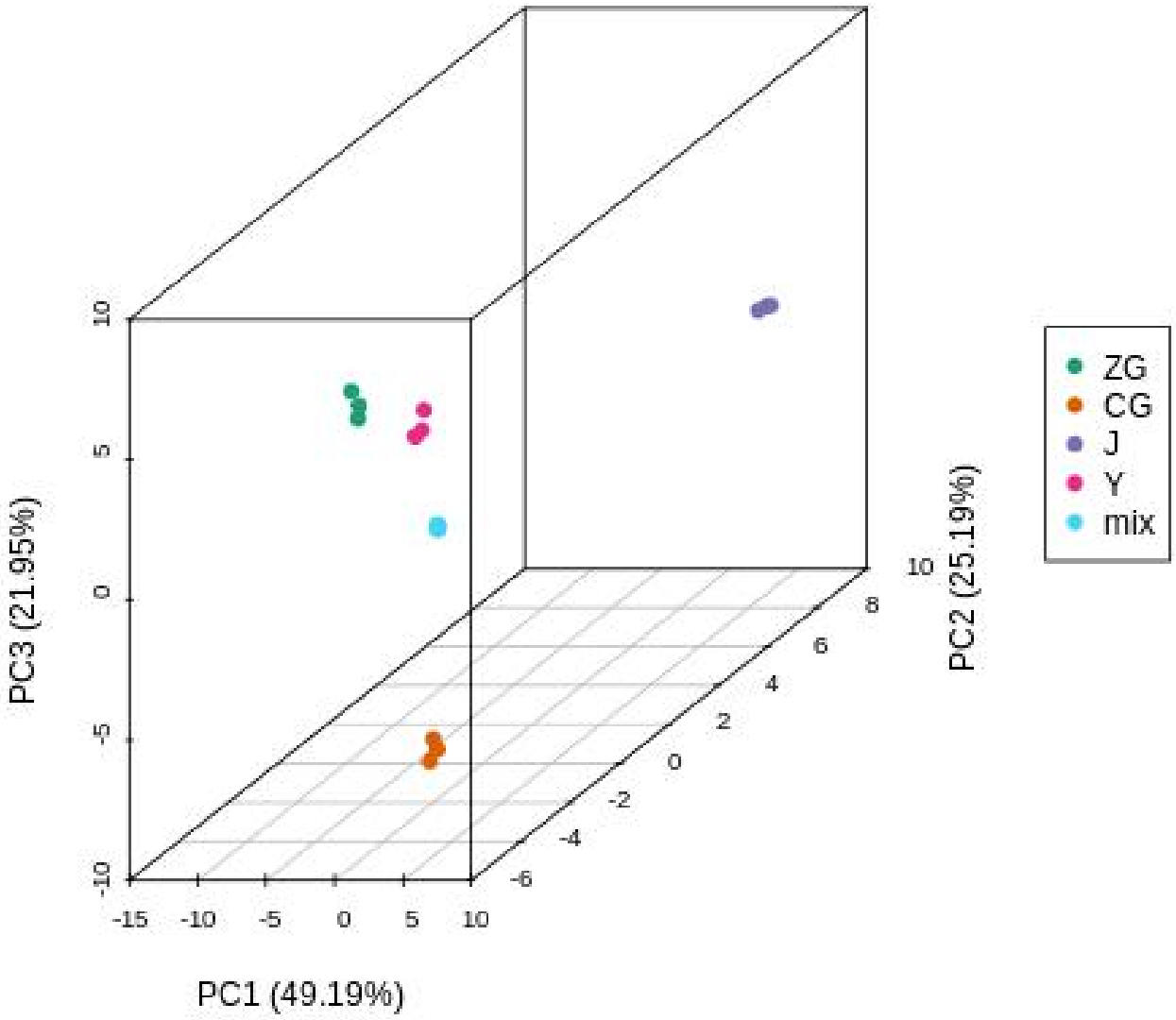

Figure S17. Principal component analysis of metabolic profiles in four tissues of *A. vilmorinianum*. Y: Leaf, J: Stem, ZG: Primary root, CG: Lateral root.

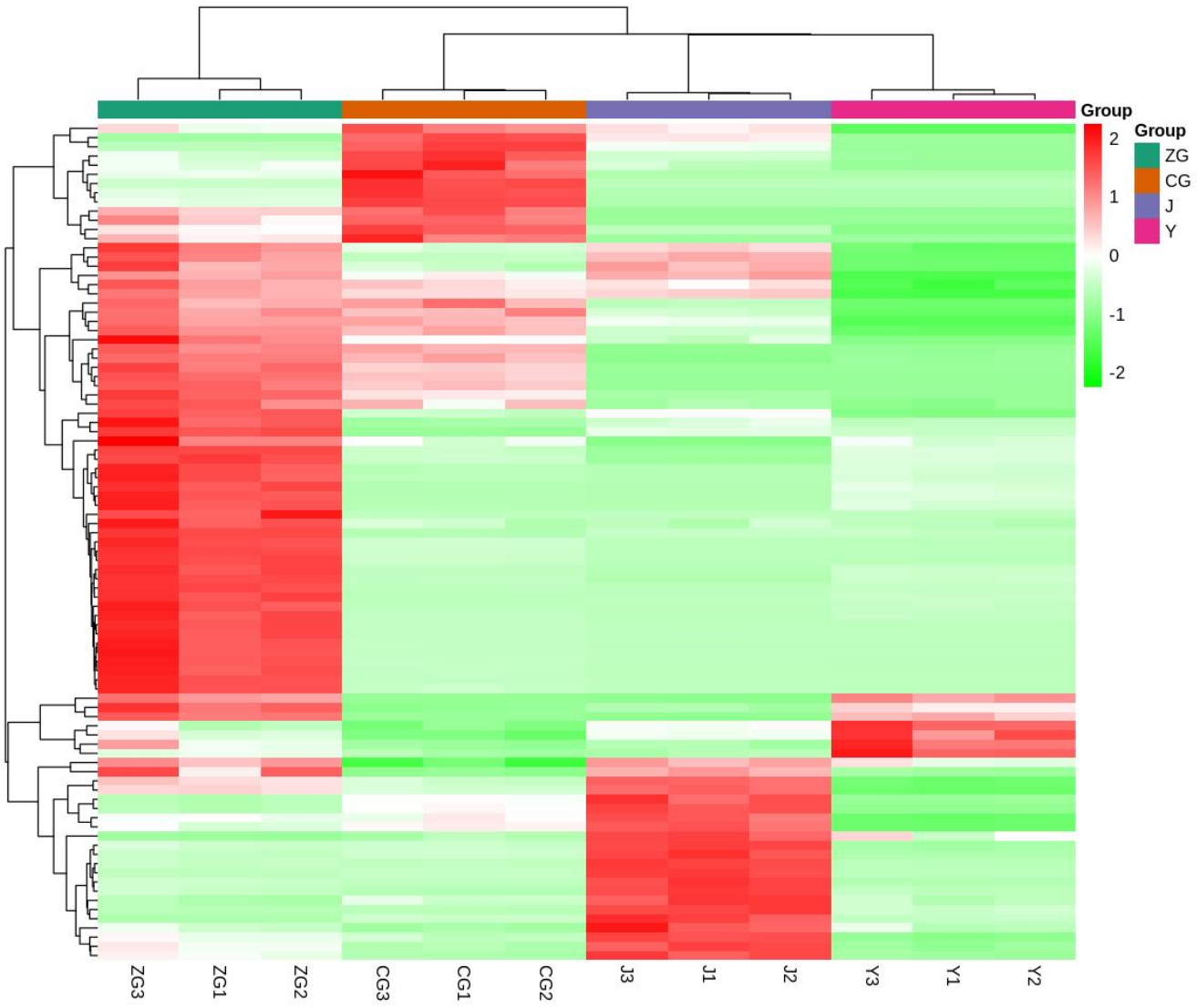

Figure S18. Heatmap of all metabolic compounds in different tissues of *A. vilmorinianum.*

Y: Leaf, J: Stem, ZG: Primary root, CG: Lateral root.

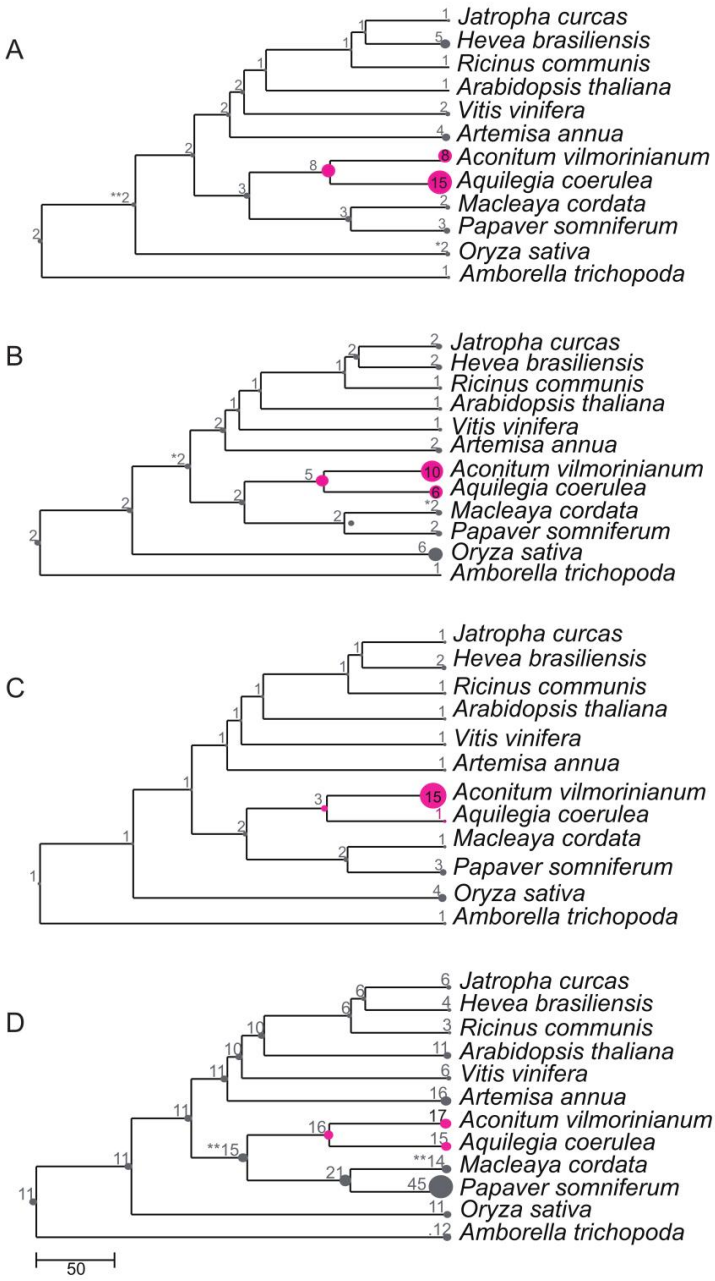

Figure S19. Expansion and contraction of orthologous involving diterpene alkaloids in *A. vilmorinianum*. A: CPSs in the OG0000968 homology. B: KSs in the OG0001660 homologrou. C: The KOXs in the OG0002270 homology. D: BAHD gene family in the OG0000043 homology.

Supporting data:

Data S1 GO functional enrichment analysis of specific expanded gene families in *A. vilmorinianum.*

Data S2 GO function enrichment analysis of clade-specific gene families.

Data S3 Summary of pathway gene annotation and expression in *A. vilmorinianum*.

Data S4 Summary of the candidate genes of 38 transcription factor families in *A. vilmorinianum*.

Data S5 Correlation analysis of transcription factors that are highly associated with diterpenoid alkaloid biosynthetic genes.

Data S6 The yield of diterpenoids in different tissues of *A. vilmorinianum*.

Data S7 Correlation analysis of genes and metabolites involved in diterpenoid alkaloid biosynthetic pathway in tissues.

Data S8 Summary of the other plant TPS genes involved in the TPS phylogenetic tree.

Data S9 Biochemically characterized aminotransferase genes in phylogenetic analysis.

Data S10 Biochemically characterized BAHD acyltransferases included in phylogenetic analysis.

Data S11 Summary of the other plant methyltransferases involved in the phylogenetic tree.

References

1 Griffiths-Jones, S. *et al.* Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121-124 (2005).

2 Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA genes in genomic sequences. *Methods Mol. Biol.* **1962**, 1-14 (2019).

3 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12**, 656-664 (2002).

4 Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988-995 (2004).

5 Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).

6 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421 (2009).

7 Benaglia, T., Chauveau, D., Hunter, D. R. & Young, D. S. Mixtools: An R package for analyzing finite mixture models. *J. Stat. Softw.* **32**, 1-29 (2009).

8 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

9 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **12**, 323 (2011).

10 Rozewicki, J., Li, S., Amada, K. M., Standley, D. M. & Katoh, K. MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Res.* **47**, W5-W10 (2019).

11 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).

12 Wang, H. *et al.* PCPD: Plant cytochrome P450 database and web-based tools for structural construction and ligand docking. *Synth. Syst. Biotechnol.* **6**, 102-109 (2021).

13 Shen, Q. *et al.* The genome of *Artemisia annua* provides insight into the evolution of Asteraceae family and artemisinin biosynthesis. *Mol. Plant* **11**, 776-788 (2018).

14 Liu, X. *et al.* The genome of medicinal plant *Macleaya cordata* provides new insights into benzylisoquinoline alkaloids metabolism. *Mol. Plant* **10**, 975-989 (2017).

15 Guo, L. *et al.* The opium poppy genome and morphinan production. *Science* **362**, 343-347 (2018).

16 Lamesch, P. *et al.* The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202-1210 (2012).

17 Zhang, L. *et al.* Global analysis of gene expression profiles in physic nut (*Jatropha curcas* L.) seedlings exposed to salt stress. *PloS one* **9**, e97878 (2014).

18 Tang, C. *et al.* The rubber tree genome reveals new insights into rubber production and species adaptation. *Nat. Plants* **2**, 16073 (2016).

19 Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463-467 (2007).

20 Chan, A. P. *et al.* Draft genome sequence of the oilseed species *Ricinus communis*. *Nat. Biotechnol.* **28**, 951-956 (2010).

21 Ouyang, S. *et al.* The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.* **35**, D883-887 (2007).
22 Albert, V. A. *et al.* The *Amborella* genome and the evolution of flowering plants. *Science* **342**, 1241089 (2013).
23 Liu, Y. et al. Analysis of the Coptis chinensis genome reveals the diversification of protoberberine-type alkaloids. Nat Commun. 12, 3276 (2021).
24 Wan, T. *et al*. A genome for gnetophytes and early evolution of seed plants. *Nat Plants* **4**, 82–89 (2018).
25 Sun, Y. *et al*. Genome sequencing of the endangered *Kingdonia uniflora* (Circaeasteraceae, Ranunculales) reveals potential mechanisms of evolutionary specialization. *iScience*. **23**(5):101124 (2020).
26 Lang, D. *et al*. The *Physcomitrella patens* chromosome-scale assembly reveals moss genome structure and evolution. *Plant J*. **93**(3):515-533 (2018).
27 Arias, T. *et al.* Genomic and transcriptomic resources for candidate gene discovery in the Ranunculids. *Appl Plant Sci*. **9**(1):e11407 (2021).
28 Wan, T. *et al*. The *Welwitschia* genome reveals a unique biology underpinning extreme longevity in deserts. *Nat Commun.* **12**: 4247 (2021).
29 Hirooka, S. *et al*. Acidophilic green algal genome provides insights into adaptation to an acidic environment. *Proc Natl Acad Sci USA*. **114**(39): E8304-E8313 (2017).
30 Yamamoto, K. et al. Three genomes in the algal genus *Volvox* reveal the fate of a haploid sex-determining region after a transition to homothallism. *Proc Natl Acad Sci USA*, **118**(21): e2100712118(2021).
31 Banks, J.A. et al.The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science* **332**(6032):960-963 (2011).
32 Pryszcz, L. P. et al. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res*, **44**:e113(2016).