

A Chromosome-Level Genome Assembly Reveals triterpenoid biosynthesis in *Clinopodium chinense*

Guohui Li

West Anhui University

Mengda Wang

West Anhui University

Muhammad Aamir Manzoor

Hainan University

Haiyu Wang

West Anhui University

Junyi Cheng

Sichuan University

Muhammad Arif

Sakarya University of Applied Sciences

Yunpeng Cao

Chinese Academy of Sciences

Jun Dai

West Anhui University

Cheng Song

lanniao812329218@163.com

West Anhui University

Research Article

Keywords: Comparative genomics, Genome assembly, *C. chinense*, Triterpenoids

Posted Date: September 12th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-7299302/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Plant Biology on November 29th, 2025. See the published version at <https://doi.org/10.1186/s12870-025-07841-8>.

Abstract

Background *Clinopodium chinense* is an important medicinal plant belonging to the Lamiaceae. The desiccated roots of *C. chinense* exhibit a variety of pharmacological properties and are utilized in traditional Chinese medicine.

Results We present the first chromosome-level genome assembly of *C. chinense*, comprising 20 pseudochromosomes with an aggregate size of 0.61 Gb and 45,466 protein-coding genes. The analysis of genome evolution indicated that two recent bursts of long terminal repeats (LTRs) significantly increased the size of the *C. chinense* genome. Additionally, numerous large-scale chromosomal rearrangements have been identified between the genomes of *C. chinense* and *Thymus quinquecostatus* genomes. Through comparative genomics studies, it was found that a recent whole-genome duplication event unique to Labiatae plants has resulted in a notable expansion of gene families related to the biosynthesis of triterpenoids and flavonoids in *C. chinense*. Subsequently, we identified several putative key genes responsible for triterpenoid biosynthesis. The results of our study offer new perspectives on the biosynthesis of triterpenoids and flavonoids, potentially advancing future investigations into the genetic and medicinal properties of *C. chinense*.

Conclusions Our research outcomes offer new perspectives on the biosynthesis of triterpenoids and flavonoids, and may aid subsequent studies on the genetic properties and medicinal uses of *C. chinense*.

Background

Clinopodium chinense (Benth.) O. Kuntze, also known by its Chinese name “Feng Lun Cai”, is classified within the genus *Clinopodium* of the Lamiaceae family. This perennial herb is prevalent across the eastern, northeastern, and southwestern regions of China [1]. In China, the aerial parts of *C. chinense*—commonly known as “duanxueliu”—have long been used in traditional folk medicine, and they are employed to treat various ailments, including hematuria, influenza, and allergic dermatitis [2]. Due to the long growth cycle, polyploid nature of the genome, and ambiguity of the phenotype, the breeding of *C. chinense* is still in its early stages. There is also a phenomenon of unstable germplasm resources in agricultural applications.

The reference genome offers insights into functional genes, evolutionary processes, and breeding practices. However, at present, only a small number of medicinal plants have been sequenced assembled into high-quality genomes [3–5]. The innovation of sequencing technology makes it possible to assemble higher quality medicinal plant genomes [6, 7]. However, due to the polyploidy of many medicinal plants, current genomic research is more difficult compared to conventional crops and horticultural plants. In earlier research, fluorescence in situ hybridization (FISH) analyses on ginseng have indicated that it is an allotetraploid [4, 8]. This type of polyploid is typically formed through the hybridization of two or more distinct diploid species followed by chromosome doubling. While genome doubling events contribute positively to boosting species diversity and environmental adaptability, this

conclusion holds particularly in the initial phases of allopolyploid formation. To coordinate and coexist with the genomes of various evolved diploid ancestral species within a single nucleus, polyploid genomes may undergo gene loss, differentiation, or even chromosome reduction, thereby acquiring a more diploid-like condition.

Previous studies have indicated that allopolyploids exhibit a subgenome advantage, wherein one subgenome typically has more genes, higher gene expression, and lower differentiation [9–11]. For example, allelic recombination occurring in the subgenome of polyploid cotton (*Gossypium hirsutum*) plays a role in its ecological adaptation and the evolution of fibers [12]. Rapid genomic changes and homologous inhibition leading to diploidization have also been observed in the subgenome of wheat (*Triticum aestivum*) [13]. In addition, the differences between the subgenomes of ginseng (*Panax ginseng*) may lead to the functional divergency and evolution of ginsenoside biosynthetic genes [4]. Therefore, elucidating the differences between the subgenomes of *C. chinense* will also contribute to functional and evolutionary research.

Like other herbs, *C. chinense* also has complex metabolites that are considered effective compounds, among which triterpenoid saponins are recognized as the most important type. Previous studies have found that *C. chinense* contain abundant flavonoids, diterpenes, and triterpenes with have anti-inflammatory, antibacterial, and procoagulant activities [14–16]. It is generally believed that the presence of secondary metabolites triterpenoid saponins in *C. chinense* gives them significant pharmacological effects, including hemostatic, anti-tumor, and hypoglycemic activities [17]. However, it is difficult to directly extract sufficient amounts of triterpenoid saponins directly from *C. chinense*. Nevertheless, the biosynthesis pathway of triterpenoid saponins has not been well characterized in various medicinal plants. Thus, conducting genome sequencing research on species like *C. chinense* could greatly enhance our comprehension of these pathways.

Triterpenoid saponins constitute a category of specialized metabolites with diverse structures, found in plants [18] as well as marine invertebrates such as sea cucumbers [19] and sponges [20]. Their biosynthetic processes involve the isoprenoid pathway, where isopentenyl pyrophosphate (IPP) serves as the precursor for all isoprenoids [21]. IPP can be produced through either the mevalonate (MVA) pathway or the 2-C-methyl-D-erythritol-4-phosphate (MEP) pathway [10]. The biosynthesis of triterpenoid saponins consists of three primary stages. Initially, isopentenyl pyrophosphate (IPP) is transformed into farnesyl pyrophosphate (FPP) through the action of geranyl-diphosphate synthase (GPPS) and farnesyl-diphosphate synthase (FPPS) [22]. In the second stage, 2,3-oxidosqualene undergoes cyclization catalyzed by 2,3-oxidosqualene cyclases—such as beta-amyrin synthase (β -AS) and lupeol synthase (LS)—resulting in the formation of various compounds like beta-amyrin and lupeol [23]. The modification stage involves a series of modifications such as oxidation, displacement, and glycosylation of terpenoid skeletons, which are carried out by specific cytochrome P450 monooxygenase (CYPs) and UDP-glycosyltransferases (UGTs) to generate triterpenoid saponins [24].

With the continuous development of genome sequencing technology, advancements in long-read sequencing techniques (e.g., HiFi sequencing) and high-throughput/resolution chromosome conformation capture (Hi-C) methods have made it possible to generate chromosome-sized genome sets in various plants. In present work, we assembled a chromosomal-level genome of *C. chinense* for traditional Chinese medicine, where two haplotypes fully represent it. We conducted a deeper investigation into the differences in triterpenoid saponin synthesis genes across subgenomes and identified gene clusters linked to triterpenoid saponin synthesis. In addition, several high-quality genome sequences of Labiatae plants published in recent years have further enhanced our understanding of the evolution and biology of Labiatae plants. The genomic resources and analyses presented in this article establish a crucial foundation for comprehending the evolutionary dynamics of Labiatae plants, while also advancing genetic research and breeding efforts for *C. chinense*.

Results

Genome sequencing, assembly, and annotation

In this study, *C. chinense* is a diploid ($2n = 2x = 38$; Fig. 1A, 1B, 1C; Table 1), with an estimated genome size of 0.61 Gb. K-mer analysis generated a total of 47,441,502,660 k-mers with a length of 19 and a peak depth of 77.96. The genome size of *C. chinense* is 608.5 Mb, with a heterozygosity of 0.89% and a duplication ratio of 54.64% (Table S1; Fig. S1). After pruning and quality control, we obtained 22.73 Gb of HiFi reads and 113.3 Gb of Hi-C paired reads (Table S2; Fig. S2). The assembled genome comprised 24 contigs with a contig N50 of 37.15 Mb and a GC content of 38.61%. The size of the assembled genome aligns closely with the projected genome size derived from k-mer analysis and flow cytometry (Table S3). The quality of *C. chinense* genome assembly in this study (N50 = 37.15 Mb) was significantly higher than that of recently published genome assemblies of other herbal plants, such as *Camptotheca acuminata* (N50 = 1.47 Mb) [25], *Morinda officinalis* (N50 = 4.21 Mb) [26], and *Senna tora* (N50 = 4.03 Mb) [25]. In addition, we further utilized Hi-C assembly information and found that these overlapping groups were further anchored to 20 pseudo-chromosomes with only one gap, covering approximately 99.12% of the assembly sequence. Genes are not evenly distributed across chromosomes; instead, a greater concentration is found at the chromosomal ends (Fig. 1D, 1E; Table S4). The length range of pseudo chromosomes is 7.21 to 44.32 Mb, with an average length of 30.16 Mb (Table S4). Furthermore, we used BUSCO software (version 5.2.2, parameter: -evalue 1e-05) [27] to evaluate of the *C. chinense* genome annotation. The results showed that approximately 96.8% of complete embryonic plant genes were identified in the assembled genome (Table 1). The BUSCO evaluation typically exceeds 90%, signifying robust genome annotation and underscoring the superior quality of the assembled genome.

Table 1
Statistics of *C. chinense* genome assembly and annotation.

Feature	Statistic
Assembled genome size (Gb)	0.61
GC content (%)	38.61
Contig number	128
Contig N50 (Mb)	34.33
Scaffold number	124
N50 (Mb)	37.15
N90 (Mb)	20.13
Minimum len (Mb)	0.02
Maximum len (Mb)	44.32
Chromosome numbers	20
Average coding sequence length (bp)	1250.95
Average exons per gene	5.51
Average exon length (bp)	337.69
Repeats in genome (%)	63.85
BUSCO (complete) (%)	96.84

The transcription element (TE) annotation of the reference genome indicated that *C. chinense* contained 61.45% repetitive sequences (Table S5), which was significantly lower than the repeat sequences found in other medicinal plant genomes. The ginseng genome comprises 83.17% repeat sequences [4], while the genome of *Dendrobium huoshanense* genome exhibits a repeat sequence value of 74.92% [5]. We further analyzed and found that long terminal repeat (LTR) sequences are the most abundant in repeat sequences, accounting for 42.42% of the *C. chinense* genome. Among them, LTR/Gypsy and LTR/Copia accounted for 19.23% and 18.86% of the genome, respectively (Table S5).

To predict protein-coding genes, we integrated ab initio, homology-based, and transcriptome-based prediction approaches (Fig. S3). Across the *C. chinense* genome, a total of 45,466 protein-coding genes were identified, with each gene having an average length of 4097.38 bp and containing 5.51 exons on average (Fig. S4; Table S6). In addition, we have conducted in-depth research and discovered a group of non-coding RNAs (Table S7).

Phylogenomic and evolution of the *C. chinense*

We performed a comparative genomic analysis of *C. chinense* with 15 other plant genomes to investigate its evolutionary history. We conducted a thorough analysis of the statistical results related to gene family identification. This analysis revealed that a total of 68,482 orthologous gene families were identified across all species, comprising 517,286 genes. The total count of single-copy gene families was 48, while the gene families common to all species amounted to 7,132, comprising 226,662 genes. The distinctive gene family of this research species comprises 3451 members, encompassing 5212 genes (Table S8; Fig. S5). To elucidate the origin of *C. chinense*, we conducted a phylogenetic tree analysis involving 15 species. Phylogenetic analysis of 48 single-copy gene families indicates that *C. chinense* is most closely linked to the *Thymus quinquecostatus* within the Labiatae family, with a divergence time of approximately 14.0 million years ago (Fig. 2A).

Whole genome duplication (WGD), or ancient polyploidization, refers to the process of genomic duplication and subsequent doubling. To investigate the WGD events that occurred during the evolution of the *C. chinense*, we studied the distribution of Ks among the selected species (Fig. 2B). The Ks value for *C. chinense*-*T. quinquecostatus* (CcTq) was 1.60. Our research indicated that members of the Labiatae family, including *C. chinense* and *S. baicalensis*, exhibit notable ancient peaks around Ks = 1.0. Furthermore, when incorporating *J. sambac* into the study, we observed that this peak emerged before the divergence peak between *C. chinense* and *J. sambac*, indicating that *C. chinense* underwent the same gamma WGD event as *J. sambac* and *S. baicalensis*.

Furthermore, evolutionary analysis of gene families showed that 6,353 gene families have expanded, accounting for 35.51% of all gene families, while 619 gene families have contracted (accounting for 3.51%). Notably, 21,736 amplified gene families and 697 contracted gene families showed statistical significance ($P < 0.05$) in *C. chinense* (Table S9). To improve our understanding of the evolutionary history of *C. chinense*, we conducted genomic collinearity research involving *C. chinense*, *T. quinquecostatus*, and *S. baicalensis* (Fig. 2C). Further analysis of different types of duplicated genes revealed that the recent WGD event specifically occurring in Labiatae contributed significantly to the expansion of these gene families.

Gene family enrichment analysis

To understand their biological functions, we conducted KEGG and GO analyses (Fig. 3). GO analysis emphasizes that significantly expanded gene families are enriched in multiple processes, such as UDP-glycosyltransferase activity, regulation of terpenoid biosynthetic process, structural constituent of cytoskeleton, regulation of coumarin biosynthetic process, and response to high light intensity (Fig. 3A; Table S10). KEGG analysis revealed that the majority of expanded genes were concentrated in pathways such as monoterpenoid biosynthesis, ubiquitin-mediated proteolysis, the biosynthesis of diverse plant secondary metabolites, ABC transporters, and plant hormone signal transduction (Fig. 3B; Table S11). However, further research has revealed that the contracted gene families were associated with GO terms linked to polysaccharide binding, oligopeptide transmembrane transporter activity, lignin catabolic process, response to auxin, lipid transport, etc. (Fig. 3C; Table S12). Furthermore, our KEGG pathway

analysis of contraction genes indicated their involvement in photosynthesis, oxidative phosphorylation, protein processing in the endoplasmic reticulum, and glutathione metabolism (Fig. 3D; Table S13). Consequently, based on the aforementioned gene family enrichment analysis, we put forward that the expansion and contraction of these genes may strengthen the adaptability of *C. chinense* in complex environments, enhance its stress tolerance and growth regulation capabilities, and ultimately ensure its survival under abiotic stress.

Co-analysis of transcriptomics and metabolomics

The use of single-omics data analysis to study complex biological processes and biological network regulation has certain limitations. The integration of multi-omics joint analysis is capable of offsetting data-related issues in single-omics data analysis, such as those arising from data loss and noise. Therefore, a combined analysis of the transcriptome and metabolome was conducted to study the biological functions of differentially expressed genes and metabolites in three different tissues of *C. chinense*: roots, stems, and leaves.. In this study, a combined analysis of transcriptome and metabolome was conducted to investigate the differential substances between different tissues of *C. chinense*. It is hoped that through this method, more comprehensive and accurate research results can be obtained, providing strong support for the development of related fields.

The joint analysis results indicated that there were 1271, 2834, and 2013 differentially expressed chemicals annotated in 136, 141, and 138 KEGG pathways, respectively, among the three comparison groups of RvsS, RvsL, and SvsL in various tissues of *C. chinense* (Fig. 4). Among them, RvsL had the most differentially expressed substances, indicating more significant differences between root and leaf tissues. The analysis of the top 20 pathways with the lowest P-values in the KEGG enrichment results showed that the category with the highest number of annotated pathways was metabolic, indicating that differentially expressed genes and metabolites in the *C. chinense* had a significant impact on plant metabolism. There were 13, 9, and 13 metabolic pathways in the three control groups, respectively. In the RvsS group, differential substances were most significantly enriched in the ribosome pathway; The differential substances in the RvsL and SvsL groups were most significantly enriched in the biosynthesis of secondary metabolites pathway, indicating that the differential substances between different tissues involved in the regulation of plant secondary metabolic pathways are mainly involved in the *C. chinense*.

Moreover, the differentially expressed genes across the three groups showed enrichment in two terpenoid biosynthesis-related pathways, specifically those involved in terpenoid skeleton biosynthesis, sesquiterpene biosynthesis, and triterpenoid biosynthesis. This suggests that the differentially expressed genes across various tissues of *C. chinense* exert a regulatory influence on terpenoid production in *C. chinense*. Considering that triterpenoid components are important active constituents in *C. chinense*, we will do extensive research on the metabolic pathways and genes related to terpenoid synthesis below.

Identification of key genes for triterpenoid saponins biosynthesis in *C. chinense*

Earlier research on the chemical components of *C. chinense* has shown that its rich triterpenoid saponins, which have been verified in clinical settings, are the primary bioactive substances [28]. Combining transcriptomic and metabolomic analysis of three different tissues undergoing *C. chinense*, we inferred and proposed the biosynthetic pathways of triterpenoid saponins and their derivatives in *C. chinense*. Terpenes are mainly synthesized through two pathways: the MVA pathway and the MEP pathway. Multiple key enzymes are involved in this process and play an important regulatory role in terpenoid synthesis (Fig. 5).

Drawing on the KEGG enrichment outcomes, our analysis centered on differentially expressed genes (DEGs) within two terpenoid biosynthesis-related pathways: Terpenoid backbone biosynthesis and Sesquiterpenoid and triterpenoid biosynthesis. We also conducted in-depth exploration and analysis of enzyme genes involved in terpenoid biosynthesis in three different tissues with *C. chinense*. In the RvsS group, there are 13 differentially expressed genes (DEGs) annotated to the Terpenoid backbone biosynthetic pathway, while 7 DEGs are annotated to the Sesquiterpenoid and triterpenoid biosynthesis pathway. Among these DEGs, we identified key enzyme genes involved in terpenoid synthesis pathways, including 3 *SEQ*, 2 *GPPS*, 2 *HMGS*, 2 *HMGR*, 1 *MVD*, 1 *DXS*, 1 *IDI*, and 1 *SS* gene. In addition, there are a total of 45 DEGs annotated to the Terpenoid backbone biosynthetic pathway and 19 DEGs annotated to the RvsL group Sesquiterpenoid and triterpenoid biosynthesis pathway. Among these DEGs, we identified key enzyme genes involved in terpenoid synthesis pathways, including 7 *SEQ*, 6 *GPPS*, 5 β -*AS*, 4 *HMGR*, 4 *PMK*, 3 *MVD*, 3 *DXS*, 3 *SS*, 2 *HMGS*, 2 *AACT*, 2 *HDR*, 2 *MVK*, 2 *IDI*, 1 *CMK*, 1 *DXR*, 1 *MCS*, and 1 *FPPS* gene. A total of 35 differentially expressed genes (DEGs) are annotated to the terpenoid backbone biosynthetic pathway, while 14 DEGs are assigned to the sesquiterpenoid and triterpenoid biosynthesis pathway belonging to the SvsL group. Among these DEGs, we identified key enzyme genes involved in terpenoid synthesis pathways, including 6 *SEQ*, 5 *HMGR*, 4 *GPPS*, 3 *MVD*, 3 *SS*, 3 β -*AS*, 2 *MVK*, 2 *AACT*, 2 *HMGS*, 2 *DXR*, 2 *IDI*, 1 *FPPS*, 1 *PMK*, 1 *CMK*, 1 *DXS*, and 1 *HDR* gene (Table. 2).

Table 2
Statistics on the number of genes for key enzymes of the terpenoid synthesis pathway.

Gene name	abbreviation	groups		
		RvsS	RvsL	SvsL
acetyl-CoA C-acetyltransferase	AACT	0	2	2
Hydroxymethylglutaryl-CoA synthase	HMGS	2	2	2
Hydroxymethylglutaryl-CoA reductase	HMGR	2	4	5
Mevalonate Kinase	MVK	0	2	2
Phosphomevalonate Kinase	PMK	0	4	1
Mevalonate pyrophosphate decarboxylase	MVD	1	3	3
1-deoxy-D-xylulose 5-phosphate synthase	DXS	1	3	1
1-deoxy-D-xylulose 5-phosphate reductoisomerase	DXR	0	1	2
4-diphosphocytidyl-2-C-methyl-D-erythritol kinase	CMK	0	1	1
2-C-methyl-d-erythritol 2,4-cyclodiphosphate synthase	MCS	0	1	
4-Hydroxy-3-methylbut-2-en-1-yl diphosphate reductase	HDR	0	2	1
Isopentenyl diphosphate isomerase	IDI	1	2	2
Farnesyl diphosphate synthase	FPPS	0	1	1
Squalene synthase	SS	1	3	3
Squalene monooxygenase	SQE	3	7	6
β -Amyrin synthase	β -AS	0	5	3
Geranylgeranyl pyrophosphate synthase	GPPS	2	6	4

The biosynthesis of triterpenoid saponin involves over 20 catalytic steps, featuring essential enzymes such as squalene synthase (SS) for scaffold formation and 2,3-oxidosqualene cyclase (OSC), alongside modifying enzymes including cytochrome P450s (CYP450) and glycosyltransferases (UGT) [29]. Furthermore, the functional enrichment analysis revealed that these DEGs were mainly enriched in "cytochrome P450" and "transporters", which may be involved in the synthesis, transport, and storage of active ingredients and stress responses in *C. chinense*. To elucidate which members of the CYP and UGT gene families may be involved in the synthesis of triterpenoid saponins in *C. chinense*. We identified 105 CYP and 21 UGT genes in the *C. chinense* genome. By prior functional predictions, we evaluated 14 CYP members and 10 UGT members potentially implicated in triterpenoid saponin biosynthesis. We additionally performed an evolutionary study on them alongside CYP and UGT members with established roles. Based on the identification of 36 CYP enzymes implicated in triterpenoid biosynthesis, we

developed a phylogenetic tree utilizing CcCYP derived from a blood cut source alongside these enzymes. We additionally used CYP101A1 from *C. chinense* as an outgroup, which is strongly associated with CYPs in plants. CYP101A1 from prokaryotes serves as an outgroup that efficiently anchors the base of the phylogenetic tree, facilitating a more precise examination of the branching direction and evolutionary distance among various CYP subfamilies. The phylogenetic tree indicated that CcCYP, which was related to triterpenoid biosynthesis, primarily clusters within the CYP85 and CYP71 subfamilies (Fig. 6A). Specifically, CcCYP707A1-7, belonging to the CYP85 family is on the same branch as GuCYP88D6, which has C-11 oxidase function. We speculated that it may be responsible for catalyzing the oxidative modification of the 11th carbon atom of triterpenoid molecules, which is an essential step in the formation of triterpenoid structures. The clustering of CcCYP26A1 with C-16 α oxidase MICYP87D16 suggests that its function may be related to catalyzing the oxidation reaction of the 16th α configuration carbon atom (such as introducing oxygen-containing groups such as hydroxyl and carbonyl groups). Meanwhile, the C-23 oxidase AtCYP71A16 is adjacent to CcCYP26A1, while CcCYP75B1A1-2 and CcCYP84 are similar to LjCYP71D353 responsible for C-20 and C-28 oxidation. CcCYP93B1-2 is also speculated to function as a C-24 oxidase potentially.

Based on the discovery of 36 glycosyltransferases involved in triterpenoid biosynthesis, we have constructed a phylogenetic tree of CcUGT with these enzymes, which are involved in *C. chinense*. As shown in the Fig. 6B, except for CcUGT72E1-2, the functions of CcUGT73C1-3 may be similar to those of oleanolic acid 3-O-glucuronosyltransferase.

Discussion

In traditional Chinese medicine, *C. chinense* is regarded as a significant medicinal plant, boasting a wealth of acknowledged bioactive components such as flavonoids, triterpenoids, polysaccharides, and amino acids [15, 30]. The analysis of whole genome sequences provides a foundation for studying the gene functions, phylogenetics, and gene evolution of *C. chinense* and many other medicinal plants. The genome of *C. chinense* has the characteristics of high heterozygosity (0.89%) and a large number of repetitive sequences (61.45%), making its assembly challenging. Moreover, medicinal plants have a relatively short domestication history, which often leads to disrupted genetic backgrounds and complicates genome mining efforts. To address this, we integrated Illumina and PacBio sequencing with high-throughput Hi-C technology to produce a high-quality *C. chinense* genome, featuring a scaffold N50 of 37.15 Mb and a contig N50 of 34.33 Mb (Table 1). Additionally, we constructed a 0.61 Gb telomere-to-telomere reference genome with a BUSCO completeness of 96.84% (Table 1). EGMA analysis further confirmed that the assembled genome fully covered 99.97% of the core eukaryotic genes. These results emphasize that compared to medicinal plant genomes sequenced recently, the integrity and quality of the *C. chinense* genome are higher. This reference genome aids in clarifying genetic variations within the saponin biosynthesis pathway throughout the evolutionary process of the *C. chinense* genome, offering guidance for the further domestication of ginseng to enhance its medicinal effectiveness.

Through phylogenomic analysis, we put forward a framework for the evolutionary process of *C. chinense*. Ks analyses showed that *C. chinense* has not experienced an WGD event after differentiation. The proliferation of different LTR-RT lineages is an important and essential factor that causes differences in plant genome size. For example, the *Tat* family is a key driver behind genome amplification in *Camellia sinensis* [31], while the *Ogre* family plays a major role in *Pisum sativum* [32]. Additionally, due to insertions, translocations, and deletions of LTR-RTs occurring during this process, incomplete autonomous LTRs now make up a substantial part of the genome. Despite the significant role LTR-RTs play in shaping genome size diversity, many aspects of them—including their origin, expression patterns, insertion specificity, evolutionary trajectories, and potential impacts on genetic and epigenetic regulation of genes—remain largely uninvestigated [33, 34].

Triterpenoids and flavonoids stand as the two primary bioactive components in *C. chinense*. Among them, triterpenoids exhibit a range of beneficial properties, including anti-inflammatory, hypotensive, and anti-aging effects, while also providing positive impacts on myocardial ischemia injury and chronic kidney disease [35]. We have tentatively verified the substantial expansion of gene families associated with the production of flavonoids and triterpenoids in AMM via comparative genomic analysis. In this study, based on the combined analysis of transcriptome and metabolomics, we identified 53 and 72 candidate genes in the biosynthesis pathways of triterpenoids and flavonoids, respectively, through homology search and functional annotation in *C. chinense*. Further analysis revealed that tandem replication may play a key role in gene amplification in the synthesis pathways of flavonoids and triterpenoid saponins, such as CYP450 and UGT genes, which are critical nodes controlling metabolic flow [36]. We hypothesized that an elevation in the copy numbers of these genes could facilitate the accumulation of active substances in *C. chinense*.

Triterpenoids are a group of natural compounds with diverse structures. They are formed when triterpenoid saponins condense with one or more sugars and other chemical groups, with tetracyclic triterpenoids and pentacyclic triterpenoids being the most common types [37]. As a result, across various tissues of *C. chinense*, differentially expressed genes (DEGs) associated with terpenoid biosynthesis were most abundant in the RvsL group, followed by the SvsL group, and least prevalent in the RvsS group, highlighting notable disparities in gene expression between root and leaf tissues and DEGs. DEGs predominantly localize within the terpenoid backbone biosynthesis pathway, indicating their significant regulatory function in terpenoid synthesis in *C. chinense*, thereby establishing a basis for the investigation of terpenoid synthesis in this species.

Conclusion

In summary, we have constructed a chromosome-level genome of *C. chinense*, establishing a basis for further research on *C. chinense* and serving as a significant resource for the detailed study of several other medicinal plants. This study identifies potential genes associated with flavonoid and triterpene production in *C. chinense*, establishing a basis for future genetic enhancement, while genomic evolutionary analysis offers novel insights into the evolutionary history of orchid species. The

investigation of the *C. chinense* genome is anticipated to elucidate its lineage selection and enhance the breeding of therapeutic plants for future generations.

Materials and Methods

Plant materials

The *C. chinense* samples were collected in August 2021 from an orchard of Taipingfan Township, Huoshan County, Lu'an, Anhui Province, China, for the purpose of sequencing analysis. High-quality DNA extracted from young roots was used to construct multiple genomic sequencing libraries. For Illumina sequencing, short-insert fragments are utilized, while long fragments are employed in SMRT sequencing.

Genome sequencing

In this study, genome sequencing was conducted via Nanopore ultra-long read, PacBio HiFi, and Hi-C sequencing technologies [38], with the work carried out by Wuhan Benegen Technology Co., Ltd. Specifically, ultra-long sequencing was performed using the Nanopore PromethION platform, and raw data were filtered to exclude failed reads with an average quality score below 7. Using Filtlong software (v0.2.4) (<https://github.com/rrwick/Filtlong>), and use Porecho v0.2.4 to cut the joint sequence (<https://github.com/rrwick/Porechop>) [39]. Subsequently, we carried out final filtering on the acquired readings (length < 30 kb, average reading quality score $\leq$ 90%) and used them for genome assembly. For the HiFi raw data generated by the Pacbio Revio sequencing platform, filtering was implemented using CCS v6.0.0 (<https://github.com/PacificBiosciences/ccs>). The CCS readings obtained are ready for subsequent analysis [40]. Assisted assembly and annotation are performed using Hi-C. Trim the Hi-C raw data using fastp v0.21.0 and then align it with the reference genome using HICUP v0.8.0 (<http://www.bioinformatics.babraham.ac.uk/projects/hicup>). We used Fastp v0.21.0 to filter the raw readings obtained from sequencing, ultimately removing low-quality, short-length, high N-based content, and adapter-contaminated readings. The repetitive sequences generated by PCR amplification were also eliminated.

Genome assembly

The *de novo* assembly of PacBio long reads was performed by Falconv0.3.0 ([github/PacificBiosciences/falcon](https://github.com/PacificBiosciences/falcon)) is performed. Firstly, all overlaps in the original reading were identified and error correction was performed on the reading using the overlap information. Overlap detection between generated correction reads for continuous components. Subsequently, the continuous sequences are polished using Quiver (a component of the SMRT analysis kit) to generate the primary assemblies. For the Bionano raw data, *de novo* assembly is performed in IrysView v. to produce consensus physical maps. The initial version of the reference genome assembly, referred to as hybrid scaffolds, is constructed by integrating the *de novo* primary assembly with Bionano genome imaging data via Bionano Solve V. As for the Hi-C sequencing data, it is mapped to the assembled scaffolds using

BWA v (available at github.com/lh3/bwa) [41] to obtain information regarding the order and orientation of the scaffolds. The scaffolds were anchored to pseudochromosomes by LACHESIS v (github.com/shendurelab/LACHESIS) [42].

Phylogenetic analyses

To identify paralogs and orthologs, as well as infer the species tree involving 15 plant species, OrthoFinder 2.5.5 was employed [43]. For example, *Camptotheca acuminata*, *Salvia miltiorrhiza*, and *Thymus quinquecostatus* et al. Construct a phylogenetic tree using identified single-copy genes. Functional annotation was performed on common and specific gene families by searching KEGG and GO databases [44]. We performed multiple sequence alignment on the single-copy gene family using Muscle v3.8.31 [45]. Next, the sequences were trimmed and aligned using Trimal v1.2rev59 with the parameter -gt set to 0.2. The processed data were then merged and fed into RAxML v8.2.10 for the construction of a maximum likelihood (ML) tree [46]. Lastly, CAFE v3.1 was employed to assess the expansion and contraction of gene families across each evolutionary branch. Gene families with significant expansion or contraction were identified based on a threshold of $p \leq 0.05$ [47].

Syntenic analyses

Homologous genes were identified both within and between species via BLASTP, and based on these findings, homologous blocks were defined using MCscanX [44]. Subsequently, WGDI was performed to calculate the Ks density of homologous and homologous gene pairs [48], and additionally, WGDI was used to generate syntenic point maps. Finally, SyRI was used to detect structural variations between subgenomes [49].

Gene duplication identification

To categorize the duplicated genes under conditions of severed blood flow, we employed DupGen Finder [50] v1.12, which enabled us to classify them into five distinct groups: WGD, TD, PD, TRD, and DSD. Subsequently, further GO and KEGG analyses were performed on the genes in these duplicate categories using the R package clusterProfiler [44] v4.0. KaKs_Calculator [51] is used to calculate the values of Ka (non-synonymous permutation for each non-synonymous site) and Ks (synonymous permutation for each synonymous site).

Abbreviations

IPP Isopentenyl pyrophosphate

MVA Mevalonate

MEP 2-C-methyl-D-erythritol-4-phosphate

FPP Farnesyl pyrophosphate

GPPS Geranyl-diphosphate synthase

β-AS Beta-amyrin synthase (β-AS)

LS Lupeol synthase

CYP450 CytochromeP450 monooxygenase

UGT UDP-glycosyltransferases

SS Squalene synthase

Declarations

Acknowledgements

We would like to thank the reviewers and editors for their careful reading and helpful comments on this manuscript.

Authors'contributions

Conceptualization: GHL, CS, MAM and JD; Software: GHL, MDW, JYC, MA and HYW; Writing-review & editing: GHL MAM, CS,YPC and JD; Data curation: GHL, MDW, JYC and HYW; Funding acquisition: GHL, JD and CS.

Data availability

The raw data generated in this study were available on the NCBI database through the Bioproject number PRJNA 1062362.

Funding

This study was supported by the Anhui Provincial University Research Projects (2023AH052637), Startup fund for high-level talents of West Anhui University (WGKQ2021079), Quality Engineering Project of West Anhui University (wxxxy2024011), Quality Engineering Project of Anhui Province (2024zybj032), Development of Big Data Integration and Analysis Platform for Traditional Chinese Medicine Genomics (0045025050), Anhui Innovation and Entrepreneurship Training Program for College Students (S202510376030), Key research and development project of Hainan Province (ZDYF2024SHFZ076) and Haikou Science and technology planning project (2022-008), and Guangxi “Bagui Young Talents” special fund.

Ethics approval and consent to participate

The experiments did not involve endangered or protected species. No specific permits were required for these locations/activities because *C. chinense* used in this study were obtained from an orchard in Taipingfan Township, Huoshan County, Lu'an, Anhui Province, China. All methods were carried out in accordance with relevant guidelines and regulations, under ethical approval and consent to participate.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Clinical trial number

This experiment is not applicable to clinical trials.

References

1. Li HY, Liu XC, Chen XB, Liu QZ, Liu ZL. Chemical composition and insecticidal activities of the essential oil of *Clinopodium chinense* (Benth.) Kuntze aerial parts against liposcelis bostrychophila badonnel, J Food Prot. 2015; 78 (10): 1870-1874.
2. Wang S, Ma G, Zhong M, Yu S, Xu X, Hu Y, Zhang Y, Wei H, Yang J. Triterpene saponins from *Tabellae clinopodii*. Fitoterapia. 2013; 90: 14-19.
3. Sun YQ, Shang LU, Zhu QH, Fan LJ, Guo LB. Twenty years of plant genome sequencing: achievements and challenges. Trends Plant Sci. 2022; 27: 391-401.
4. Song YT, Zhang YT, Wang X, Yu XK, Liao Y, Zhang H, Li LF, Wang YP, Liu B, Li W. Telomere-to-telomere reference genome for *Panax ginseng* highlights the evolution of saponin biosynthesis. Horticulture Research. 2024; 11: uhae107.
5. Han BX, Jing Y, Dai J, Zheng T, Gu FL, Zhao Q, Zhu FC, Chen CW, Yue Z, Chen NF A chromosome-level genome assembly of *Dendrobium Huoshanense* using long reads and Hi-C data. Genome Biol Evol. 2020; 12(12): 2486-2490.
6. Younessi-Hamzekhanlu M, Ozturk M, Jafarpour P, Mahna N. Exploitation of next generation sequencing technologies for unraveling metabolic pathways in medicinal plants: a concise review. Ind Crop Prod. 2022; 178: 114669.
7. Bielecka M, Pencakowski B, Nicoletti R. Using next-generation sequencing technology to explore genetic pathways in endophytic fungi in the syntheses of plant bioactive metabolites. Agriculture. 2022; 12: 187.
8. Choi HI, Waminal NE, Park HM, Kim NH, Choi BS, Park MY, Choi DI, Lim YP, Kwon SJ, Park BS, Kim HH, Yang TJ. Major repeat components covering one-third of the ginseng (*Panax ginseng* CA Meyer) genome and evidence for allotetraploidy. Plant J. 2014; 77: 906-16.
9. Nie SA, Zhao SW, Shi TL, Zhao W, Zhang RG, Tian XC, Guo JF, Yan XM, Bao YT, Li ZC, Kong L, Ma HY, Chen ZY, Liu H, El-Kassaby YA, Porth IG, Yang FS, Mao JF. Gapless genome assembly of azalea and multi-omics investigation into divergence between two species with distinct flower color. Hortic Res. 2023; 10: uhac241.

10. Cheng F, Wu J, Cai X, Liang JLI, Freeling MA, Wang XW. Gene retention, fractionation and subgenome differences in polyploid plants. *Nat Plants*. 2018; 4: 258-68.
11. Wang MJ, Tu LL, Lin M, Lin ZX, Wang PC, Yang QY, Ye ZX, Shen C, Li JY, Zhang L, Zhou XL, Nie XH, Li ZH, Guo K, Ma YZ, Huang C, Jin SX, Zhu LF, Yang XY, Min L, Yuan DJ, Zhang QH, Lindsey K, Zhang XL. Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nat Genet*. 2017; 49: 579-87.
12. Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin DC, Llewellyn D, Showmaker KC, Shu SQ, Udall J, Yoo MJ, Byers R, Chen W, Faigenboim AD, Duke MV, Gong L, Grimwood J, Grover C, Grupp K, Hu GJ, Lee TH, Li JP, Lin LF, Liu T, Schmutz J. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature*. 2012; 492: 423-427.
13. Levy AA, Feldman M. Evolution and origin of bread wheat. *Plant Cell*. 2022; 34: 2549-67.
14. Lili Li, Qi Huang, Xianchun Duan, Lan Han, Daiyin Peng. Protective effect of *Clinopodium chinense* (Benth.) O. Kuntze against abnormal uterine bleeding in female rats. *J Pharma Sci*. 2020; 02(004): 1347-8613.
15. Zhu YD, Hong JY, Bao F, Xing N, Wang LT, Sun ZH, Luo Y, Jiang H, Xu XD, Zhu NL, Wu HF, Sun GB, Yang JS. Triterpenoid saponins from *Clinopodium chinense* (Benth.) O. Kuntze and their biological activity. *Arch Pharm Res*. 2018; 41(12): 1117e1130.
16. Zhang HJ, Chen RC, Sun GB, Yang LP, Zhu YD, Xu XD, Sun XB. Protective effects of total flavonoids from *Clinopodium chinense* (Benth.) O. Kuntze on myocardial injury in vivo and in vitro via regulation of Akt/Nrf2/HO-1 pathway. *Phytomedicine*. 2018; 40: 88e97.
17. Zhu YD, Hong JY, Bao FD, Xing N, Wang LT, Sun ZH, Luo Y, Jiang H, Xu XD, Zhu NL, Wu HF, Sun GB, Yang JS. Triterpenoid saponins from *Clinopodium chinense* (Benth.) O. Kuntze and their biological activity. *Arch Pharm Res*. 2018; 41: 1117-1130.
18. Augustin JM, Kuzina V, Andersen SB, Bak S. Molecular activities, biosynthesis and evolution of triterpenoid saponins. *Phytochemistry*. 2011; 72: 435-457.
19. Mondol MAM, Shin HJ, Rahman MA, Islam MT. Sea cucumber glycosides: chemical structures, producing species and important biological properties. *Mar Drugs*. 2017; 15: 317.
20. Kalinin VI, Ivanchina NV, Krasokhin VB, Makarieva TN, Stonik VA. Glycosides from Marine Sponges (Porifera, Demospongiae): Structures, Taxonomical Distribution, Biological Activities and Biological Roles. *Mar Drugs*. 2012; 10: 1671-1710.
21. Zhao CL, Cui XM, Chen YP, Liang Q. Key enzymes of triterpenoid saponin biosynthesis and the induction of their activities and gene expressions in plants. *Nat Prod Commun*. 2010; 5: 1147-1158.
22. Haralampidis K, Trojanowska M, Osbourn AE. Biosynthesis of Triterpenoid Saponins in Plants. *Adv. Biochem. Eng Biotechnol*. 2002; 75: 31.
23. Vincken JP, Heng L, de Groot A, Gruppen H. Saponins, classification and occurrence in the plant kingdom. *Phytochemistry*. 2007; 68: 275-297.
24. Sawai S, Saito K. Triterpenoid biosynthesis and engineering in plants. *Front Plant Sci*. 2011; 2: 25.

25. Kang SH, Pandey RP, Lee CM, Sim JS, Jeong JT, Choi BS, Jung M, Ginzburg D, Zhao K, Won SY, Oh TJ, Yu Y, Kim NH, Lee OR, Lee TH, Bashyal P, Kim TS, Lee WH, Hawkins C, Kim CK, Kim JS, Ahn BO, Rhee SY, Sohng JK. Genome enabled discovery of anthraquinone biosynthesis in *Senna tora*. Nat Communication. 2021; 11: 5875.
26. Wang J, Xu S, Mei Y, Cai S, Gu Y, Sun M, Liang Z, Xiao Y, Zhang M, Yang S. A high quality genome assembly of *Morinda cinalis*, a famous native southern herb in the lingnan region of southern China. Hortic Res. 2021; 8: 135.
27. Manni M, Berkeley MR, Seppey M, Simao FA, Zdobnov EM. "BUSCO update: Novel and streamlined work flows along with broader and deeper phylogenetic coverage of eukaryotic, prokaryotic, and viral genomes," arXiv preprint arXiv. 2021; 2106: 11799.
28. Shi YY, Zhang SX, Peng DY, Wang CK, Zhao D, Ma KL, Wu JW, Huang LQ. Transcriptome Analysis of *Clinopodium chinense* (Benth.) O. Kuntze and Identification of Genes Involved in Triterpenoid Saponin Biosynthesis. Int J Mol Sci. 2019; 20: 2643.
29. Hou MQ, Wang RF, Zhao SJ, Wang ZT. Ginsenosides in Panax genus and their biosynthesis. Acta Pharm Sin B. 2021; 11: 1813-34.
30. Gao YL, Wang YZ, Wang KZ, Zhu J, Li GS, Tian JW, Li CM, Wang ZH, Li J, Leed AW, Guo CH. Acute and a 28-day repeated-dose toxicity study of total flavonoids from *Clinopodium chinense* (Benth.) O. Ktze in mice and rats. Regul Toxicol Pharm. 2017; 91: 117-123.
31. Zhang QJ, Li W, Li K, Nan H, Shi C, Zhang Y, Dai ZY, Lin YL, Yang XL, Tong Y. et al. The chromosome-level reference genome of tea tree unveils recent bursts of non-autonomous LTR retrotransposons in driving genome size evolution. Mol Plant. 2022; 13: 935-938.
32. Kreplak J, Madoui MA, Ca'pal P, Nova'k P, Labadie K, Aubert G, Bayer PE, Gali KK, Syme RA, Main D. A reference genome for pea provides insight into legume genome evolution. Nat Genet. 2019; 51: 1411-1422.
33. Zhao M, and Ma J. Co-evolution of plant LTR-retrotransposons and their host genomes. Protein Cell. 2013; 4: 493-501.
34. Yi C, Fang T, Su H, Duan SF, Ma RR, Wang P, Wu L, Sun WB, Hu QC, Zhao MX, Sun LJ, Dong XH. A reference-grade genome assembly for Astragalus mongholicus and insights into the biosynthesis and high accumulation of triterpenoids and flavonoids in its roots. Plant Communications. 2022; 4: 100469.
35. Zhang CH, Yang X, Wei JR, Chen NMH, Xu JP, Bi YQ, Yang M, Gong X, Li ZY, Ren K. Ethnopharmacology, phytochemistry, pharmacology, toxicology and clinical applications of Radix Astragali. Chin J Integr Med. 2021; 27: 229-240.
36. Seki H, Tamura K, and Muranaka T. P450s and UGTs: key players in the structural diversity of triterpenoid saponins. Plant Cell Physiol. 2015; 56: 1463-1471.
37. Thimmappa R, Geisler K, Louveau T, O'Maille P, and Osbourn A. Triterpene biosynthesis in plants. Annu Rev Plant Biol. 2014; 65: 225-257.

38. Gong L, Wong CH, Idol J, Ngan CY, Wei CL. Ultra-long Read Sequencing for Whole 959 Genomic DNA Analysis. *Journal of visualized experiments: JoVE*. 2019.
39. Bonenfant Q, Noé L, Touzet H. Porechop_ABL: discovering unknown adapters in Oxford 961 Nanopore Technology sequencing reads for downstream trimming. *Bioinformatics* 962 *advances*. 2023; 3: vbac085.
40. Tamura KT, Teranishi YG, Ueda SY, Suzuki HY, Kawano NK, Yoshimatsu K, Saito KK, Kawahara NB, Muranaka TY, Seki H. Cytochrome P450 Monooxygenase CYP716A141 891 is a Unique β -Amyrin C-16 β Oxidase Involved in Triterpenoid Saponin Biosynthesis in 892 *Platycodon grandiflorus*. *Plant cell physiol*. 2017; 58: 874-884.
41. Li HX, Wu S, Lin RX, Xiao YR, Morotti ALM, Wang Y, Galilee MT, Qin HW, Huang T, Zhao Y, Zhou X, Yang J, Zhao Q, Kanellis AK, Martin C, Tatsis EC. The genomes of medicinal skullcaps reveal the polyphyletic origins of clerodane diterpene biosynthesis in the family Lamiaceae. *Mol Plant*. 16: 549-570.
42. Wolff J, Rabbani L, Gilsbach R, Richard G, Manke T, Backofen R, Gruning BA . Galaxy HiCExplorer 3: a web server for 976 reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and 977 visualization. *Nucleic Acids Res*. 2020; 48: W177-184.
43. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. 2019; 20:1-14.
44. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T. ClusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovations*. 2021; 2: 100141.
45. Yu HW, Wang HX, Liang X, Liu J, Jiang C, Chi XL, Zhi NN, PSu P, Zha LP, Gui SY. Telomere-to-telomere gap-free genome assembly provides genetic insight into the triterpenoid saponins biosynthesis in *Platycodon grandiflorus*. *Hortic Res*. 2025; 12: uhaf030
46. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. AL: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)*. 2009; 25: 1972-3.
47. Mamta, Shikha S, Sangeeta K, Poonam P, Aasim M Gopal S, Ram KSa. High-quality haplotype-resolved chromosome assembly provides evolutionary insights and targeted steviol glycosides (SGs) biosynthesis in *Stevia rebaudiana* Bertoni. *Plant Biotechnol J*. 2024; 22: 3262-3277.
48. Sun PC, Jiao BB, Yang YZ, Shan LX, Li T, Li XN, Xi ZX, Wang XY, Liu JQ. WGDl: a user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Mol Plant*. 2022; 15: 1841-51.
49. Goel M, Sun HQ, Jiao WB, Schneeberger KB. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol*. 2019; 20: 1-13.
50. Qiao X, Li QH, Yin H, Qin KJ, Li LT, Wang RZ, Zhang SL, Paterson AH. Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. *Genome Biol*. 2019; 20: 1-23.
51. Zhang Z. KaKs_calculator 3.0: calculating selective pressure on coding and non-coding sequences. *Genom Proteom Bioinf*. 2022; 20: 536-40.

Figures

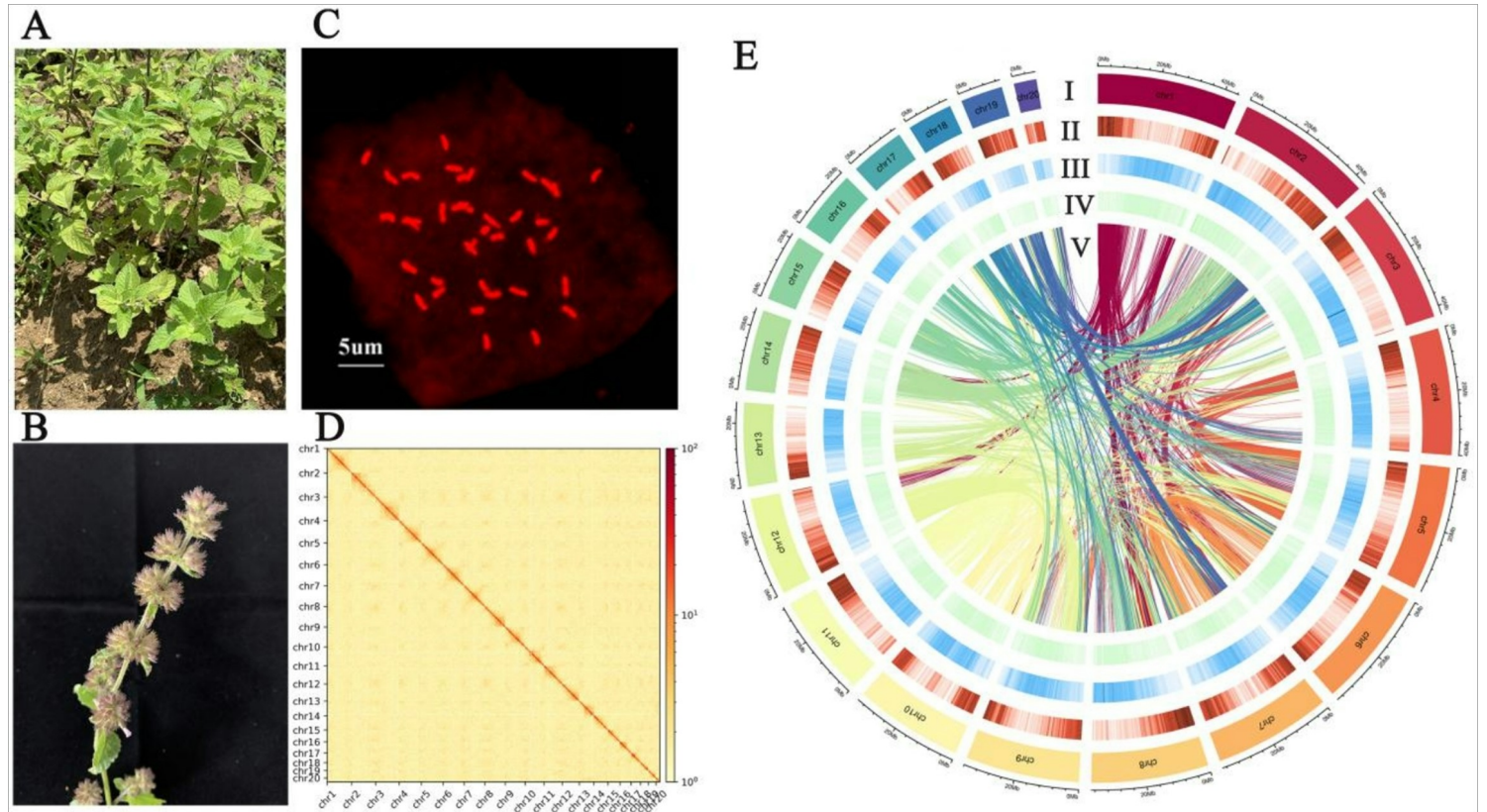


Figure 1

The habitat, phenotypes, karyotypes, and genomic features of *C. chinense*.

A: Cultivation and growth of *C. chinense* in the field. **B:** Morphology of the leaf and stem of *C. chinense*. **C:** Fluorescence staining of *C. chinense* chromosome with DAPI ($2n=2x=38$). **D:** Hi-C map of *C. chinense* of the draft genome assembly. **E:** Genomic features of the genome assembly. (I) The 20 assembled *C. chinense* chromosomes. (II) The gene count along the genome. (III) The repetitive sequence density along the genome. (IV) The GC content along the genome. (V) Syntenic relationships among different chromosomes of *C. chinense*.

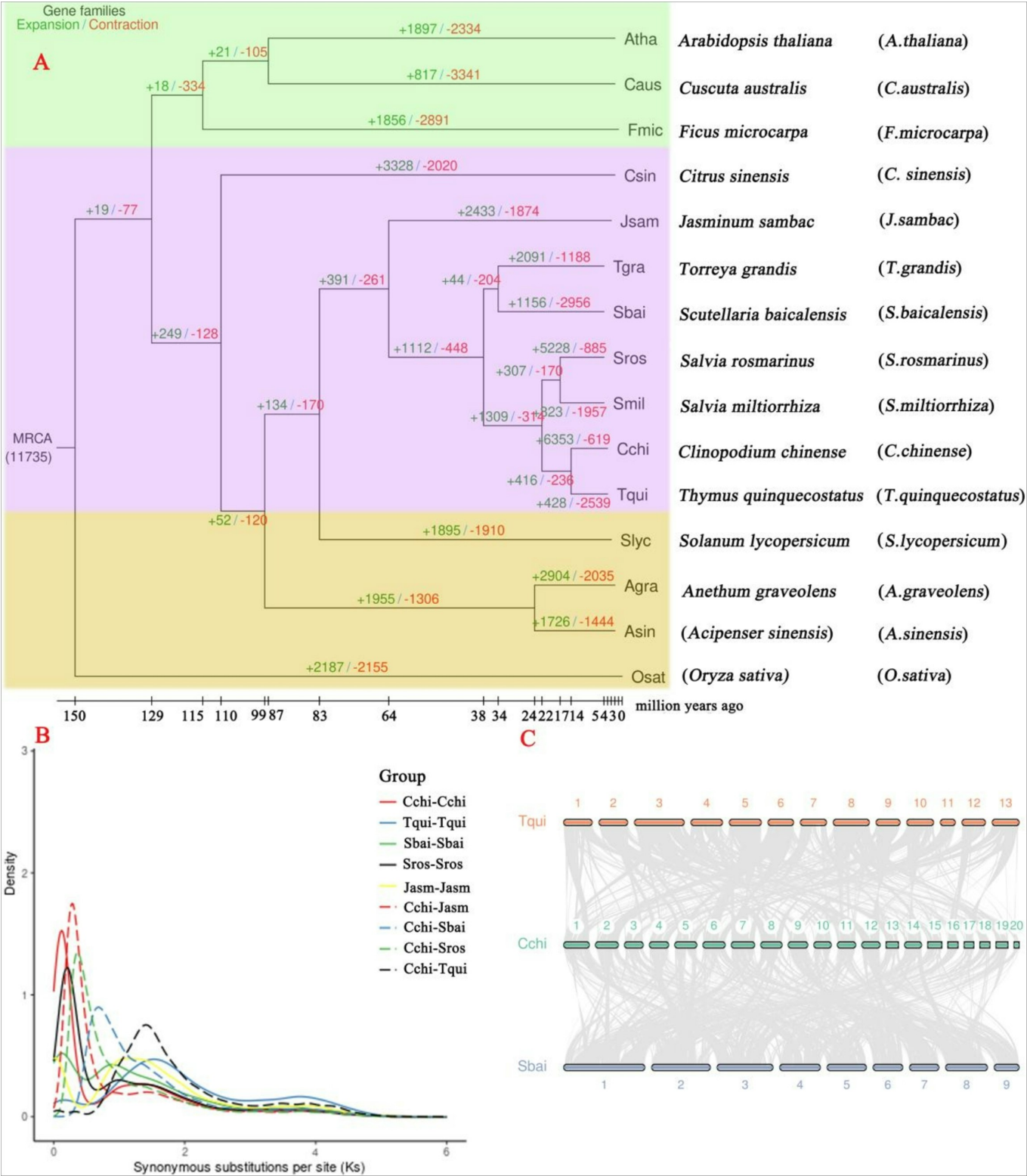


Figure 2

Comparative genomic and evolutionary analysis of *C. chinense*.

A:Phylogenetic tree of 15 plant species and the expansion/contraction of gene families. The numbers on the nodes indicate estimated divergence times (million years ago [Mya]), and the blue bars show the error range. The numbers in green and red indicate the expanded and contracted gene families in the

lineage, respectively. The divergence time of speciation events are superimposed on the tree below. The background color represents the tribe to which the species belongs. All the nodes have 100% bootstrap support. MRCA, most recent common ancestor; WGD, whole genome duplication. **B:** Ks distribution between *C. chinense*, *T. quinquecostatus*, and *S. baicalensis*. *Astragalus membranaceus* var. *mongholicus* is abbreviated as *A. mongholicus*. The peak location is denoted by dotted lines. Ks, synonymous substitution. **C:** Synteny between *C. chinense* (20 chromosomes), *T. quinquecostatus* (13 chromosomes), and *S. baicalensis* (9 chromosomes).

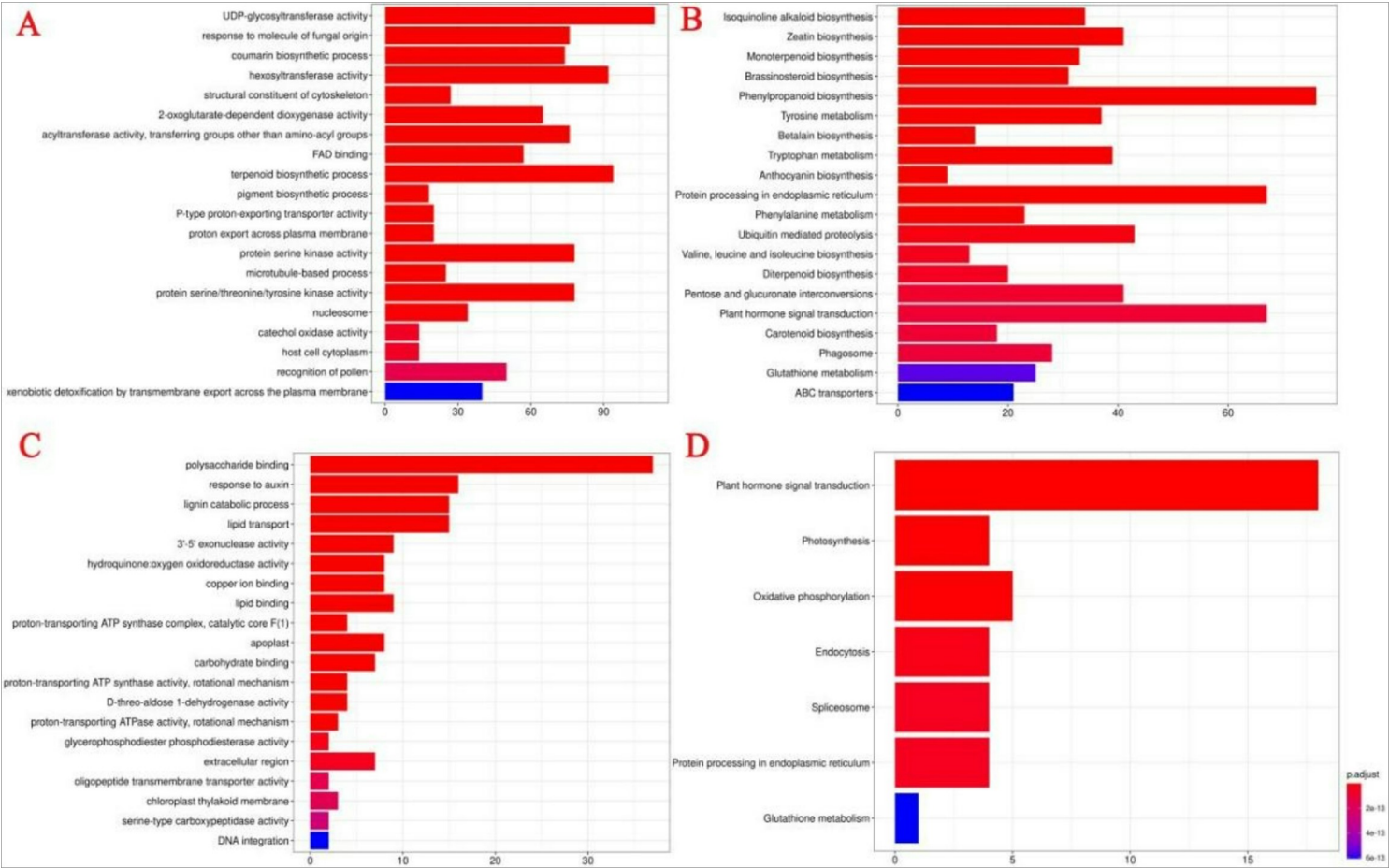


Figure 3

Gene family GO and KEGG enrichment analysis.

- A:** Gene ontology (GO) enrichment analysis conducted on the genes with significant expansion.
- B:** KEGG enrichment analysis performed on the genes that are significantly expanded.
- C:** GO enrichment analysis of the significantly expanded genes.
- D:** KEGG enrichment analysis of the contracted genes.

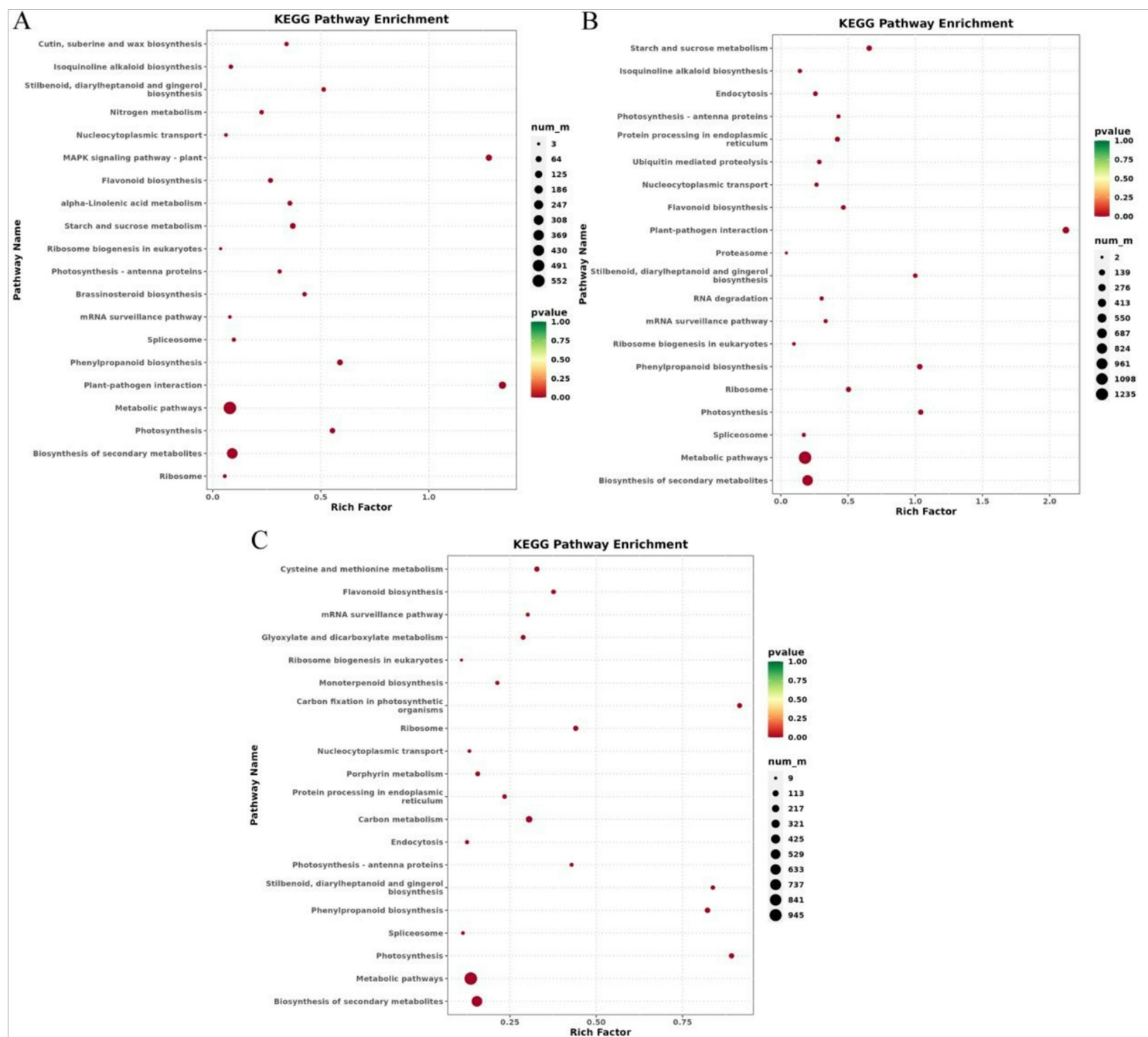


Figure 4

Co-analysis of transcriptional metabolism of KEGG enrichment maps.

A: KEGG enrichment map of RvsS. **B:** KEGG enrichment map of RvsL. **C:** KEGG enrichment map of SvsL.

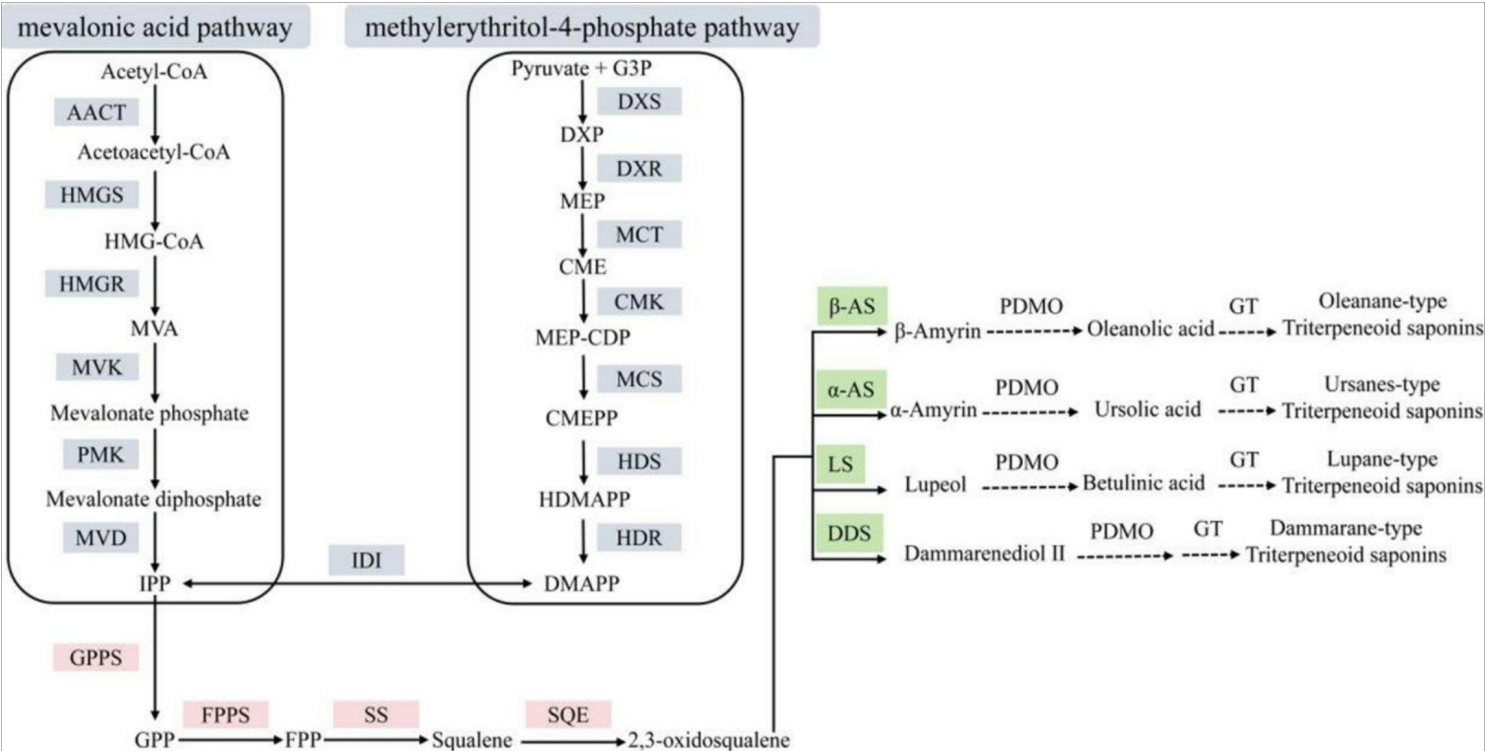


Figure 5

Terpenoid biosynthetic pathways.

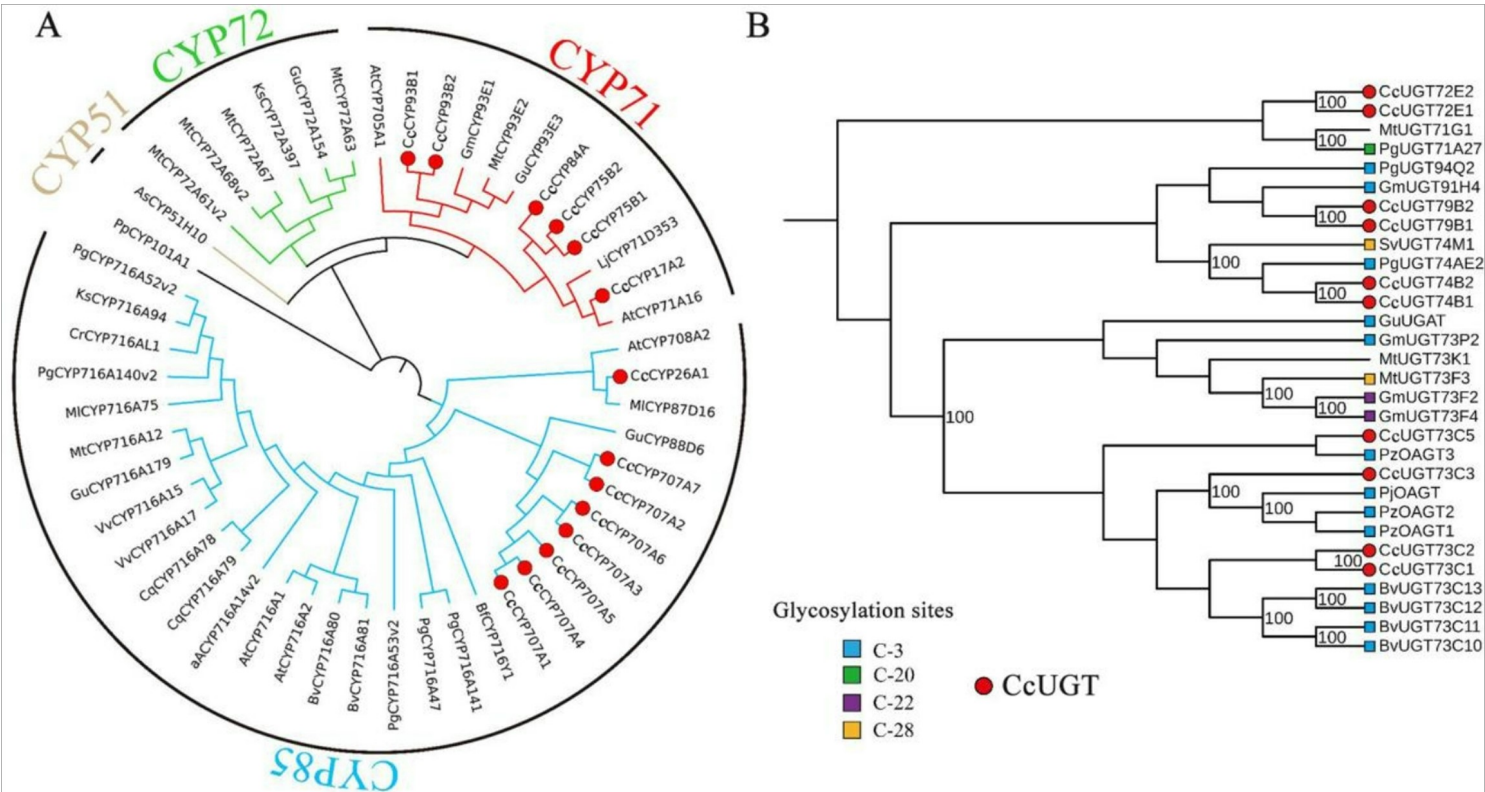


Figure 6

Phylogeny of *CYP* and *UGT* gene family families identified in the *C. chinense* genome.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformationlegends.docx](#)
- [Supplementarymateria.rar](#)