Diploid Chromosome-level Reference Genome and Population Genomic Analyses Provide Insights into Gypenoside Biosynthesis and Demographic Evolution of Gynostemma pentaphyllum (Cucurbitaceae)

Running title: Genomics and demographic evolution of Gynostemma pentaphyllum

Xiao Zhang¹, Yuhe Zhao¹, Yixuan Kou², Xiaodan Chen³¹³, Jia Yang¹, Hao Zhang¹⁴, Zhe Zhao¹, Yuemei Zhao⁵, Guifang Zhao¹⁺, Zhonghu Li¹⁺

1. Key Laboratory of Resource Biology and Biotechnology in Western China (Ministry of Education), College of Life Sciences, Northwest University, Xi’an, Shaanxi, 710069, China
2. Laboratory of Subtropical Biodiversity, Jiangxi Agricultural University, Nanchang, 330045, China
3. College of Life Sciences, Shanxi Normal University, Taiyuan, Shanxi, 030012, China
4. College of Life Sciences, Sun Yat-sen University, Guangzhou, Guangdong, 510275, China
5. School of Biological Sciences, Guizhou Education University, Guiyang, Guizhou, 550018, China

⁺. Corresponding authors: Zhonghu Li, Tel./Fax number: +86-029-88302411
Guifang Zhao, Tel./Fax number: +86-029-88305264

Xiao Zhang, zhxiao@nwu.edu.cn
Yuhe Zhao, zhaoyuhe@stumail.nwu.edu.cn
Yixuan Kou, yixuankou@jxau.edu.cn
Xiaodan Chen, chenxiaodan@stumail.nwu.edu.cn
Jia Yang, yjhgx@stumail.nwu.edu.cn
Hao Zhang, zhangh623@mail2.sysu.edu.cn
Zhe Zhao, zhaozhe@stumail.nwu.edu.cn
Yuemei Zhao, zhaoyuemei@gznc.edu.cn
Guifang Zhao, gfzhao@nwu.edu.cn
Zhonghu Li, lizhonghu@nwu.edu.cn

© The Author(s) 2022. Published by Oxford University Press. All rights reserved. This is an Open Access article distributed under the terms of the Creative Commons Attribution License https://creativecommons.org/licenses/by/4.0/, which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Abstract

Gynostemma pentaphyllum (Thunb.) Makino is a perennial creeping herbaceous plant in the family Cucurbitaceae, which has great medicinal value and commercial potential, but urgent conservation efforts are needed due to the gradual decreases and fragmented distribution of its wild populations. Here, we report the high-quality diploid chromosome-level genome of G. pentaphyllum obtained using a combination of next-generation sequencing short reads, Nanopore long reads, and Hi-C sequencing technologies. The genome is anchored to 11 pseudo-chromosomes with a total size of 608.95 Mb and 26,588 predicted genes. Comparative genomic analyses indicate that G. pentaphyllum is estimated to have diverged from Momordica charantia 60.7 million years ago, with no recent whole-genome duplication event. Genomic population analyses based on genotyping-by-sequencing and ecological niche analyses indicated low genetic diversity but a strong population structure within the species, which could classify 32 G. pentaphyllum populations into three geographical groups shaped jointly by geographic and climate factors. Furthermore, comparative transcriptome analyses showed that the genes encoding enzyme involved in gypenoside biosynthesis had higher expression levels in the leaves and tendrils. Overall, the findings obtained in this study provide an effective molecular basis for further studies of demographic genetics, ecological adaption, and systematic evolution in Cucurbitaceae species, as well as contributing to molecular breeding, and the biosynthesis and biotransformation of gypenoside.

Keywords: demographic evolution, ecological niche analysis, genome assembly and annotation, gypenoside biosynthesis gene, population genomics

INTRODUCTION

Gynostemma pentaphyllum (Thunb.) Makino belongs to the family Cucurbitaceae and it is a perennial creeping herb with palmately compound leaves, dioecious flowers, and globose berries. This species is the most widely distributed plant in Gynostemma Bl., where it is mainly distributed in moist and warm mountainous forests in sub-tropical East and Southeast Asia, with a wide elevation range from 300 m to 3200 m. In China, the wild plant resources of G. pentaphyllum are highly abundant in the Yangtze River Basin and its southern areas. Polyploidization is common and complex in G. pentaphyllum natural populations and the mechanism responsible for the formation of its mixed-ploidy populations is still unclear. As a traditional Chinese medicinal plant, G. pentaphyllum is generally used due to its anti-oxidation, anti-cancer, and anti-inflammatory properties, as well as functioning in reducing blood fat and improving immunity because it contains many beneficial substances, such as saponins, flavonoids, and amino acids. Since the isolation of the first two gypenosides in 1981, more than 200 gypenosides have been isolated and identified, and about 25% have similar structures to ginsenosides. Thus, Gynostemma species are considered important resources for extracting
ginsenosides. Previous studies suggested that gypenosides and ginsenosides are both damarane-type tetracyclic triterpenoid saponins, and that gypenosides are secondary metabolites in the synthesis pathway for triterpenoids. However, further studies are still necessary, especially of the modification stage after triterpenoid saponin skeleton formation, which is an important stage that often leads to the generation of various gypenosides. Therefore, analyzing the biosynthetic pathway for gypenosides will help to further regulate the biosynthesis of gypenosides, and lay the foundations for obtaining gypenosides and their biotransformation into ginsenosides.

However, *G. pentaphyllum* has now been listed as a Grade II Key Protected Wild Plant Species by the Chinese government due to the excessive harvesting of its wild resources, which had led to gradual decreases and the fragmented distribution of *G. pentaphyllum* wild populations. Therefore, extensive investigations and studies of the population genetics of this species appear to be urgently required, which will contribute to the formulation of a strategy for conserving the population size but also genetic variations. Few previous molecular genetic studies of *G. pentaphyllum* have been conducted and most were based on molecular markers, such as Inter-simple Sequence Repeat (ISSR), Random Amplified Polymorphic DNA (RAPD), and Simple Sequence Repeats (SSR), as well as plastid and nuclear DNA fragments sequences obtained by Sanger sequencing.

Whole genome sequences can provide important resources for studying the origin and evolution of plant species, as well as genetic variations, traits for crop improvement, and biosynthesis pathways. Moreover, there are many potential advantages of using genotyping-by-sequencing (GBS) for resolving the evolutionary history and genetic structure of plant populations, and genomic selection.

Here, we report a high-quality diploid chromosome-level *G. pentaphyllum* genome obtained using next-generation sequencing (NGS), third-generation Oxford Nanopore Technologies, and Hi-C technologies. We also screened the changes in the expression levels of a series of candidate genes related to the biosynthesis pathway for gypenosides by comparative transcriptome sequencing analysis. Furthermore, *G. pentaphyllum* population genomics were investigated using genome-wide single nucleotide polymorphisms (SNPs) obtained through GBS combined with ecological niche comparison analyses to elucidate the evolution and demographic history of natural *G. pentaphyllum* populations in sub-tropical China. The results obtained in this study provide an important reference genome for further population evolution studies, as well as for the exploitation, utilization, and conservation of *G. pentaphyllum* species resources. Our findings also enrich the genomic sequences available in Cucurbitaceae to facilitate evolutionary studies of this important plant family. In addition, our findings contribute to a better understanding of the dynamic history of plant populations in East Asia.

**RESULTS**

**Genome Ploidy Evaluation**

According to flow cytometry and analysis of anatomical sections of root tips, a diploid *G. pentaphyllum* plant individual was identified (2n = 2x = 22; Figures 1d and 1e). Based on NGS short reads, the characteristics of the genome were evaluated from a total of 101.74 Gb of clean data. Results obtained by k-mer frequency analysis estimated the genome size of *G. pentaphyllum* as about 591–592 Mb, with a heterozygosis rate of 1.49–1.55% and 66.4–72.16% repeat sequences (Figure S1, Table S1), thereby indicating that the *G. pentaphyllum* genome is highly...
genomically intact (Table 1), demonstrating the high quality of the assembled genome. The characteristics of the 11 pseudo-chromosomes in *G. pentaphyllum* are shown in Figure 1c and Table S3.

**Genome Annotation**

We annotated 366.74 Mb repeats (60.23% of the genome) in the *G. pentaphyllum* genome, and 361.07 Mb (59.30% of the genome) of these repeats were identified as transposable elements (Table 1 and Table 2). The most abundant repetitive sequences were long terminal repeat (LTR) retrotransposons, which accounted for 48.28% of the genome, followed by DNA (9.15%), long interspersed nuclear elements (1.82%), and short interspersed nuclear elements (0.05%; Table 2). In total, 27,418 protein-coding genes (CDSs) were predicted in the *G. pentaphyllum* genome. The average CDS length and exon number per gene were 2,350 bp and 10, respectively (Table S5). BUSCO evaluation showed that 91.5% of the Embryophyta odb10 gene set was completely covered by the genome annotations (Table S6). In particular, 96.97% (26,588 of 27,418) of the predicted genes were functionally annotated in eight public databases (Table S7), where 17,664 genes were shared in five databases (Figure 1b). Moreover, we identified non-coding RNAs including 110 miRNAs, 631 tRNAs, 1,161 rRNAs, and 459 snRNAs (Table S8).

**Comparative and Evolutionary Genomic Analyses**

Comparative genomic analyses were performed by comparing the *G. pentaphyllum* genome with the genomes of 11 other plant species comprising nine Cucurbitaceae species, one Datiscaceae species, and one Fabaceae species (Table S9). In total, 187,702 gene families (orthogroups) comprising 333,354 genes were determined across the 12 species. In addition, we identified 2,713 *G. pentaphyllum*-specific genes in these gene families (Figure 1h and Table S10). Orthogroup gene statistics were calculated for each species, and the *G. pentaphyllum* genome contained fewer (6,664) single-copy genes than most of the species compared, except for three Cucurbita species and *Glycine max* (Figure 1h, Table S11).

Analysis of gene family expansion and contraction found that 3,247 gene families expanded and 4,598 gene families contracted, which accounted for 13.92% and 19.72% of all gene families, respectively (Figure 1g). Furthermore, 2,242 genes in all of the significantly expanded gene families (*p* < 0.05) were enriched in 85 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, such as “Plant–pathogen interaction,” “Tryptophan metabolism,” and “Isoflavonoid biosynthesis” (Figure S2a and Table S12). In the significantly contracted gene families (*p* < 0.05), 16 genes were enriched in seven KEGG pathways comprising “Benzoxazinoid biosynthesis,”
Phagosome,” “Phenylpropanoid biosynthesis,” “Tryptophan metabolism,” “Endocytosis,” “Biosynthesis of secondary metabolites,” and “Metabolic pathways” (Figure S2b and Table S13).

We constructed a phylogenetic tree based on the 12 species using 6,108 genes from 509 single-copy gene families. The tree showed that G. pentaphyllum diverged earlier than other cucurbits in the Cucurbitaceae family, where the divergence time from Momordica charantia was 60.7 million years ago (Mya), with a confidence interval of 50.2–71.1 Mya (Figure 1g). The Ks distribution of orthologs suggested that a recent whole-genome duplication (WGD) event was absent in G. pentaphyllum as in other species in the Cucurbitaceae family, whereas Cucurbita pepo and Datisca glomerata underwent a recent WGD event (Figure 1i). Speciation event times were obtained using the calculated divergence time for G. pentaphyllum – D. glomerata at 90.7 Mya. The Ks value of G. pentaphyllum – D. glomerata was 1.05, which allowed us to calculate the age of an ancient WGD (Ks = 1.674) at 144.58 Mya according to the computational formula: T = Ks/2r. In addition, complex genome collinearity was observed among G. pentaphyllum and C. lanatus, C. pepo, and C. sativus (Figure S3), thereby indicating that the genomes of these species had undergone considerable chromosomal rearrangement.

Transcriptome Sequencing and Analysis of Gypenoside Biosynthetic Pathway Genes

In total, 98.86 Gb of clean transcriptome data was obtained from 16 samples, and the average Q20 was 96.70% (Table S13). Assembly and annotation were performed by mapping clean reads to the G. pentaphyllum genome. We screened out 788 genes involved in the gypenoside biosynthetic pathway. Comparative transcriptome analyses of five tissues (after averaging the replicate samples; Figure S4) were conducted, and 58 differentially expressed genes (DEGs) were selected to assess the gene expression profiles (Figure 2 and Table S15). The results identified six genes involved in the formation stage, including one geranylgeranyl diphosphate synthase (GPPS) gene, two farnesyl diphosphate synthase (FPS) genes, two squalene synthase (SS) genes, and one squalene epoxidase (SE) gene, where the GPPS and FPS genes were significant DEGs. In the modification stage that causes diversification, the cyclization reaction of 2,3-oxidosqualene catalyzed by 2,3-oxidosqualene cyclase (OSC) gene, causes diversification, the cyclization reaction of 2,3-oxidosqualene catalyzed by 2,3-oxidosqualene cyclase (OSC) gene, two farnesyl diphosphate synthase (FPS) genes, two squalene synthase (SS) genes, and one squalene epoxidase (SE) gene, where the GPPS and FPS genes were significant DEGs. In the modification stage that causes diversification, the cyclization reaction of 2,3-oxidosqualene catalyzed by 2,3-oxidosqualene cyclase (OSC) gene, two farnesyl diphosphate synthase (FPS) genes, two squalene synthase (SS) genes, and one squalene epoxidase (SE) gene, where the GPPS and FPS genes were significant DEGs. In the modification stage that causes diversification, the cyclization reaction of 2,3-oxidosqualene catalyzed by 2,3-oxidosqualene cyclase (OSC) gene, two farnesyl diphosphate synthase (FPS) genes, two squalene synthase (SS) genes, and one squalene epoxidase (SE) gene, where the GPPS and FPS genes were significant DEGs. In the modification stage that causes diversification, the cyclization reaction of 2,3-oxidosqualene catalyzed by 2,3-oxidosqualene cyclase (OSC) gene, two farnesyl diphosphate synthase (FPS) genes, two squalene synthase (SS) genes, and one squalene epoxidase (SE) gene, where the GPPS and FPS genes were significant DEGs. In the modification stage that causes diversification, the cyclization reaction of 2,3-oxidosqualene catalyzed by 2,3-oxidosqualene cyclase (OSC) gene, two farnesyl diphosphate synthase (FPS) genes, two squalene synthase (SS) genes, and one squalene epoxidase (SE) gene, where the GPPS and FPS genes were significant DEGs. In the modification stage that causes diversification, the cyclization reaction of 2,3-oxidosqualene catalyzed by 2,3-oxidosqualene cyclase (OSC) gene, two farnesyl diphosphate synthase (FPS) genes, two squalene synthase (SS) genes, and one squalene epoxidase (SE) gene, where the GPPS and FPS genes were significant DEGs. In the modification stage that causes diversification, the cyclization reaction of 2,3-oxidosqualene catalyzed by 2,3-oxidosqualene cyclase (OSC) gene, two farnesyl diphosphate synthase (FPS) genes, two squalene synthase (SS) genes, and one squalene epoxidase (SE) gene, where the GPPS and FPS genes were significant DEGs. In the modification stage that causes diversification, the cyclization reaction of 2,3-oxidosqualene catalyzed by 2,3-oxidosqualene cyclase (OSC) gene, two farnesyl diphosphate synthase (FPS) genes, two squalene synthase (SS) genes, and one squalene epoxidase (SE) gene, where the GPPS and FPS genes were significant DEGs. In the modification stage that causes diversification, the cyclization reaction of 2,3-oxidosqualene catalyzed by 2,3-oxidosqualene cyclase (OSC) gene, two farnesyl diphosphate synthase (FPS) genes, two squalene synthase (SS) genes, and one squalene epoxidase (SE) gene, where the GPPS and FPS genes were significant DEGs.
performed based on 107 G. pentaphyllum individuals, which divided the 32 populations into three groups (Figures 3a and 3c) as the cross validation (cv) value was lowest at K = 3 (Figure 3f). In particular, the ADMIXTURE results showed that when K = 2, the 32 populations could be divided into two groups according to their geographic locations in the north and south (Group N and Group S, respectively), and when K = 3, Group N was stable but Group S was further divided into two groups designated as Group SE and Group SW, which geographically occupied the southeastern and southwestern regions, respectively (Figures 3a, 4b). Most populations were purely compositional but some were partially mixed (mosaic) in the Ta-pa Mountains–Wushan Mountains region and southeastern edge of Yunnan–Kweichow Plateau (Figure 3a). It should be noted that the AL population located on Taiwan island did not cluster with the adjacent mainland but instead it clustered together with Group SW. Analysis of molecular variance (AMOVA) detected significant genetic differentiation among the three groups (FST = 0.42, p < 0.001), but only 5.44% of the genetic variation was partitioned among the groups (Table 3).

We constructed a non-rooted branching maximum-likelihood (ML) tree to determine the phylogenetic relationships among 107 individuals and all individuals clustered into three main clades with high bootstrap values. Principal component analysis (PCA) also divided these individuals into three groups, where the three major PCs explained 24.77% (11.64%, 8.44%, and 4.69%, respectively) of the total variance (Figure 3g).

The effective population size of G. pentaphyllum determined based on pairwise sequentially Markovian coalescent (PSMC) analysis increased initially to $-2 \times 10^5$ individuals but then started to decline around 1.1 Ma (Figure 3d). The results obtained by using the PSMC approach were quite limited in terms of the sample size and predictions for recent years, so we also determined the recent demographic history of G. pentaphyllum populations with SNP data using SMC++. For recent timescales, the global trend agreed with the PSMC results. However, further more refined tests found different changes in the effective population sizes of the three groups. In particular, Group SW had undergone continuous decline whereas Group SE expanded around 12–13 thousand years ago (ka). It was interesting that the change in the effective population size of Group N followed a pattern of expand–contract–expand, where the changes occurred at approximately 100 ka and 12 ka (Figure 3e). Analysis of the effective population sizes of the three groups at 10 ka showed that Group SE (~5500 individuals) was larger than Group N (~3500 individuals) and Group SW (~2000 individuals) (Figure 3e).

**Niche Comparison Analyses**

The potential present distribution range predicted with MAXENT was consistent with the current geographic distribution of G. pentaphyllum, with area under the receiver operating characteristic curve values over 0.95 (Figure S5). Compared with the present, the distribution range in the last interglacial (LIG) period was much wider, whereas that in the last glacial maximum (LGM) period was significantly restricted, and the overall historical dynamic trend comprised initial shrinkage and then expansion. However, the main distribution center areas were quite different, where they focused in eastern and southeastern China in the LIG period, before retreating to the Qinling and Ta-Pa Mountains in central China in the LGM period, and then expanding to the present distribution pattern mainly in the west and southwest (Figure S5). Separate consideration of the three groups clearly indicated that the N and SE groups underwent dynamic changes consistent with most taxa, but Group SW was characterized by an in-situ expansion–contraction process (Figure S5).
Kernel density plots were generated to depict the frequency distributions of each group for six selected environmental variables, which showed that the three groups occupied different environments (Figure 3j–o). According to PCA-env analysis, the first two axes explained 83.11% of the total variation in the climatic conditions across the G. pentaphyllum population distributions for the three groups. The first component (PC1) and second component (PC2) explained 59.32% and 23.79% of the total variance, respectively (Figure 3h). PC1 was mainly explained by Bio 14 (temperature seasonality (STD*100)) and PC2 was explained by Bio 15 (precipitation seasonality (CV)) (Figure S6). Multiple niche plots for the 20% occurrence density and metrics of niche dynamics obtained by comparing pairs showed that Group N and Group SE shared their climate niche to a great extent, whereas Group SW had a relatively independent niche (Figure 3h–o and Table S18).

**DISCUSSION**

Cucurbitaceae is the fourth largest economically important edible botanical family in the world[31]. Many genomic studies have investigated gourds[22, 26, 32], but the early divergence of taxa in this family is still not fully understood. According to phylogenetic studies of a wide range of species, G. pentaphyllum is among the oldest taxa in the Cucurbitaceae family[31, 33] and its ploidy composition is complex[19]. Therefore, sequencing the diploid whole genome of G. pentaphyllum can help to obtain a deeper understanding of the origin and evolutionary processes for the Cucurbitaceae family, as well as providing a reference for further studies of the mechanism associated with the development of mixed polyploids G. pentaphyllum populations.

In this study, we successfully sequenced and assembled a chromosome-level high-quality diploid G. pentaphyllum genome and anchored the sequences to 11 pseudo-chromosomes, which corresponded to the karyotype (2n = 2x = 22). Compared with a previous genome assembly for this species, the contig N50, scaffold N50, and BUSCO values were all significantly improved (Table 1), and the final genome size (608.95 Mb) was larger than that of the previous assembly (582 Mb)[34]. In the Cucurbitaceae family, the genome size of G. pentaphyllum is similar to that of Sechium edule (606.42 Mb)[35] and smaller than that of B. hispida (859 Mb)[22], but much larger than most gourd genomes, such as those of C. lanatus (367.91 Mb)[36], L. siceraria (313.4 Mb)[36], and C. maxima (271.4 Mb)[37] (Table S9). In total, 26,588 (96.97% of the predicted genes) CDSs were annotated with functions in G. pentaphyllum. The number of predicted genes in G. pentaphyllum is less than those in most gourd genomes[22] excluding watermelon[38], snake gourd[59], and bottle gourd[36].

Based on comparisons with 11 related plant species, our gene family results identified 2,713 G. pentaphyllum-specific genes (Figure 1h, Table S10), where 3,247 gene families expanded and 4,598 gene families contracted (Figure 1g). The expansion and contraction of gene families are regarded as key drivers of adaptive evolution and stress tolerance in various species[40-42]. Further analyses could focus on the associated molecular evolutionary mechanisms and functional analyses of these gene families.

WGD events are of great significance for species diversification, morphological diversity, and the acquisition of new functions during evolution[42, 43]. Previous studies have demonstrated that Cucurbitaceae plants underwent four WGD events, where an early large-scale cucurbit-common tetraploidization (CCT) occurred shortly after the ancient WGTγ event (approximately 130–150 Mya) shared by all core eudicots, and three recent WGDs within some
genera in the tribe Cucurbitaeae, tribe Sicyoeae, and tribe Gomphogyneae. However, we did not detect the CCT event and no recent WGD events were found in *G. pentaphyllum*, which are consistent with previous studies based on whole genome data. These differences might have been caused by the use of different data set types (such as transcriptome or genome data) and analytical approaches. In addition, the quite low rates of preserved CCT collinear genes in cucurbit genomes might have led to the failure to detect the CCT event. The peak $Ks$ distribution (at 144.58 Mya) evaluated for *G. pentaphyllum* overlapped with the time of the ancient WGTγ event (Figure 1i). Therefore, we suggest that the complex ploidy within *G. pentaphyllum* may have been the result of adaptation to different ecological niches by this widely distributed species, which will be another key focus of our future research. The phylogenetic tree and estimated divergence time indicated that *G. pentaphyllum* diverged earlier than other cucurbits in the Cucurbitaceae family, where the divergence time from *M. charantia* was around 60.7 Mya. This result supports a previous conclusion that the genus *Gynostemma* originated in the Early Tertiary Period.

The evolution of a species is largely related to its genetic diversity. According to our results, the genetic diversity of *G. pentaphyllum* was low and similar to that determined in previous genetics research using microsatellite molecular markers. In addition, *G. pentaphyllum* is a dioecious perennial herbaceous plant that mainly reproduces asexually in a suitable environment. Long-term clonal growth will inevitably lead to decreased genetic diversity within a population and increased genetic differences between populations. Moreover, low diversity may be related to the population dynamics history of the species.

Current genetic structure patterns are produced by evolutionary and demographic processes at different temporal scales. Plant palaeoecology reconstruction provides fundamental guidance for testable phylogeographic hypotheses, but it cannot indicate the detailed population history. Ecological niche analysis can be used to predict the distributions of species under the impact of climate change and overcome the limitation due to inadequate fossil records in East Asia. Combining ecological niche analysis with molecular data can enhance our understanding of population dynamics in the past and future. In the present study, PCA and phylogenetic analysis of individuals identified three genetic groups, which were highly consistent with those obtained by ADMIXTURE clustering analysis (Figures 3c and 3g). According to our results, most of the populations exhibited very large geographical differences, and purely compositional differences might have been the result of founder or bottleneck effects. However, the three groups were partially mixed (mosaic phenomenon) in the Ta-pa Mountains–Wushan Mountains region (ES, JS, FJ, and JY populations) and the southeastern edge of Yunnan–Kweichow Plateau (BS, YN, and KM populations; Figure 3a), thereby indicating that genetic exchanges had occurred between these populations throughout their evolutionary history. We propose that these areas might have been glacial refuges or recent differentiation centers for *G. pentaphyllum* during the LGM period.

In its early demographic history, the effective population size of *G. pentaphyllum* increased initially and then started to decline around 1.1 Mya (Figure 3d). The recent changes in the three groups differed to some extents (10–100 ka, Figure 3e). According to MAXENT distribution simulations, *G. pentaphyllum* populations underwent shrinkage initially before then expanding after the LIG period (130 ka, Figure S5). At the start of the LGM period (21 ka), the Earth’s climate was extremely cold and vast areas at high latitudes were covered by continental
Thus, *G. pentaphyllum* populations retreated to the southeastern edge of the Qinghai–Tibet Plateau, Yunnan–Kweichow Plateau, and Qinling–Ta-pa Mountains areas. After the ice age, the temperature gradually warmed and herbaceous plants that preferred the humid environment gradually flourished. The *G. pentaphyllum* populations expanded from two large refuges and migrated along both sides of the rivers. The northern population in the Qinling–Ta-pa Mountains spread along the Yangtze River to the middle and lower reaches of the Yangtze River plain, whereas the southern population on the Qinghai–Tibet Plateau and Yunnan–Kweichow Plateau moved southward along the Pearl River to south China, thereby forming the current distribution pattern. In addition, our results indicated that the eastern edge of the Yunnan–Kweichow Plateau and the Ta–pa–Wushan Mountains area, which are located at the boundary of the second and third steps of topography in China, played very important roles in the formation of the genetic structure of *G. pentaphyllum* in subtropical China in a similar manner to that of the genetic structure of *Quercus fabri* in the same area. The kernel density based on six selected environmental variables and E-space analysis indicated that Group N and Group SE partly shared their climate niche, whereas no niche overlap was found for Group SW (Figure 3h–o). Ecological differentiation provides preconditions for the differentiation of lineages following initial spatial isolation and also leads to adaptation to their corresponding environmental conditions to some degree. Thus, the differences in the climatic conditions and ecological niches for the three groups over time eventually led to heterotopic differentiation of the three groups. Overall, our results indicate that geographic and climate factors shaped the genetic structure of *G. pentaphyllum*.

Gypenoside is the main active substance in *G. pentaphyllum*, and it has remarkable medicinal and nutraceutical properties. Based on recent studies of the gypenoside biosynthesis pathway, we explored the expression levels in different tissues of specific genes encoding enzymes that regulate each step in the gypenoside biosynthesis pathway. Most of the genes had higher expression level in the leaves and tendrils, as also suggested in a previous study and these findings were consistent with the detection of the highest gypenoside content in the leaves of *G. pentaphyllum*. Therefore, analyzing genes in the gypenoside biosynthesis pathway could facilitate further regulation of gypenoside biosynthesis, and provide the foundations for obtaining gypenoside and its biotransformation into ginsenosides.

In conclusion, in this study, we obtained a high-quality diploid chromosome-level whole genome sequence of *G. pentaphyllum* and compared it with the genomes of related taxa, thereby enriching the genome information available regarding the early differentiation of species in Cucurbitaceae. Our findings provide a basis for further phylogenetic studies of Cucurbitaceae, particularly understanding the mechanism associated with the formation of complex mixed polyploidy in *G. pentaphyllum*. Moreover, genome-level population genetics were investigated using a large number of populations within the distribution areas, which further improved our understanding of the population genetic diversity in *G. pentaphyllum*. By comparing the ecological niches among groups, we determined the population dynamics history of *G. pentaphyllum* and the mechanism associated with the formation of its genetic structure, thereby providing a reference for the rational utilization and development of *G. pentaphyllum* and species diversity protection. In addition, the identification of potential genes in the gypenoside synthesis pathway provides a strong theoretical basis for improving the gypenoside content through molecular breeding in the future.
MATERIALS AND METHODS

Plant Materials

A female individual *G. pentaphyllum* was collected from Qingcheng Mountain (30°56’N, 103°34’E) in Sichuan Province, China, and cultured at the Evolutionary Botany Laboratory of Northwest University. Root tips were immersed in FAA fixative solution and used for chromosomal evaluation. Young buds aged 7 days were frozen at –80°C after processing in liquid nitrogen for Hi-C sequencing. Fresh leaves aged 15 days were used for flow cytometry analysis. Healthy leaves were placed in silica gel to dry rapidly and used for total DNA extraction.

Sixteen fresh samples of five tissues (four young leaves, four stems, four tendrils, four flowers, and one fruit) were also collected from the Qingcheng Mountain population and stored at –95°C.

Healthy leaves from 107 wild individuals in 32 populations were collected from the main distribution area of *G. pentaphyllum* in China (Figure 3a and Table S13), and dried rapidly in silica gel. A handheld GPS (GarminTrex Handheld GPS; Garmin) was used to determine the latitude and longitude of each site. Voucher specimens for the samples were deposited at Northwest University (Xi’an, Shaanxi, China).

Genome Ploidy Evaluation

Due to the prevalence of polyploidy within *G. pentaphyllum* species, the chromosomal ploidy of samples should be determined first. Chromosome karyotyping was conducted using anatomical sections and flow cytometry to identify authentic diploid *G. pentaphyllum* plants. Root tips were soaked in Kano fixing fluid for 24 h, before suspending in 1.0 mol/L HCl for 15 min, dyeing with modified carbolic–basic fuchsin liquid, tabletting in the conventional manner, and observing by optical microscopy at 100× to count the chromosome numbers. Flow cytometry was conducted using a Partec CyFlow Space system by Jindi Future Biotechnology (Beijing, China).

DNA Isolation

Total DNA was extracted from dry leaves using the modified CTAB method and stored at 4°C. DNA samples for genome sequencing were sent to BGI-Shenzhen (Shenzhen, China) for NGS, Nanopore library construction, and sequencing, and the genomic DNA samples for GBS were sent to Novogene (Beijing, China).

Genome Sequencing

The short-read (PE 150 bp) NGS data were sequenced on the DNBSEQ™ platform. The total DNA was randomly fragmented and then sequenced after quality control. The raw reads were filtered using SOAPnuke v1.6.5 to remove low quality reads, adapter pollution, and PCR duplication. The Nanopore (20 kb) library was constructed for long-read sequencing on the PromethION Beta v19.05.1 platform (Oxford Nanopore Technologies, Oxford, UK). The raw reads were quality filtered with thresholds comprising: minimum read length > 5000 bp and Q value of read > 7. The high quality reads (clean reads) were then used to assemble the genome.

Genome Size Estimation

The genome size of *G. pentaphyllum* was assessed by k-mer frequency analysis using NGS short reads. After filtering using SOAPnuke v1.6.5, the optimal k-mer size was analyzed by KmerGenie. The k-mer counts were then analyzed using Jellyfish and GenomeScope was employed to fit of the frequency spectrum, before estimating the genome size, heterozygosity, and repeat sequences.
Genome Assembly

Genome assembly was conducted as described previously using Necat v0.01, Racon v1.3.3, Medaka v0.7.1, Pilon v1.23, and purge haplotigs. The Burrows–Wheeler Aligner (BWA) v0.7.8 and Samtools v1.9 were used to align filtered short fragments to the assembled genome sequence, before applying BamDeal v0.22 to obtain statistics regarding the GC distribution and depth with a window of 10 kb.

Hi-C Sequencing

The Hi-C library was prepared using chromatin from the nuclei of young buds, which were fixed with formaldehyde and extracted. High-throughput sequencing was performed for each qualified library. Raw data were also filtered using SOAPnuc v1.6.5. Valid interaction pairs were identified and extracted from the clean Hi-C data, and then aligned to the genome sketch using Juice. Based on the alignment results, clustering, sorting, and orientation were performed to assist with genome anchoring into chromosomes using 3D-dna software. The assembled chromosome-level genome was split into bins of 100 kb to construct an interaction heatmap for validation and manual correction. The completeness of the assembled genome was assessed on the Embryophyta odb10 database using BUSCO v4.0.6 with the default parameters. To assess the genomic integrity and accuracy, the short reads were also mapped back onto the assembled genome using BWA.

Genome Annotation

We employed de novo and homology-based approaches to annotate the repeat sequences. RepeatMasker v. 4.0.72 and RepeatProteinMask were used to align the transposable elements against the RepBase library (RepBase 21.01). In the de novo method, RepeatScout, Piler, RepeatModeler-2.1, and LTR_FINDER v. 1.07 were used to construct a de novo repeat sequence library, and the repeat sequences were identified using RepeatMasker against this library. In addition, tandem repeats were isolated using Tandem Repeats Finder v. 4.0.77 and LTR elements were investigated with LTR_FINDER v. 1.07. Finally, the overlapping repeat sequences obtained by using the two methods were merged together.

Ab initio, homolog, and transcriptomic evidence based strategies were used to predict the gene structure using Maker v.2.31.878 (details in Note S1). Final gene sets were counted and the numbers of genes supported by the three prediction methods were retained. Finally, the completeness of the gene set was assessed based on comparisons with the plant single-copy ortholog gene database (Embryophyta odb10) using BUSCO 30.

The gene sets obtained from gene structure annotation were searched against eight known protein databases (NCBI non-redundant protein (NR/NT), Gene Ontology (GO), KEGG, Clusters of Orthologous Groups (COG), SwissProt, InterPro, and TrEMBL) using BLAST v2.2.31 with an e-value less than 1e-5.

Comparative Genomic and Evolutionary Analyses

Comparative genomic and evolutionary analyses were performed among G. pentaphyllum and 11 related species (Table S9). All-vs-all BLAST analysis was conducted based on protein sequences from all 12 species using BLASTP v2.2.26 with an e-value of 1e-5, and OrthoMCL v2.0.9 was then used to perform gene family clustering. In total, 6,108 orthologous genes were collected from 509 single-copy gene families and aligned with Mafft. MrBayes v3.1.2 and PhyML were used to construct the phylogenetic tree based on the second bases in these gene codons (phase 1 site), and TreeBeST was employed for root determination. MCMCTREE in the
PAML v4.5\textsuperscript{84} package was used to estimate the molecular clock (substitution rate) and the divergence times among species by using *Glycine max* as an outgroup. Fossil calibration points from Timetree\textsuperscript{85} were used to inform the species divergence times.

Genome collinear blocks were identified using MCS\textsuperscript{86} with the default parameters, where each contained at least 20 and five collinear gene pairs when comparing different chromosomes in the genome and among species, respectively. Collinearity plots were generated using the JCVI program. To identify possible WGD events, we compared the genome of *G. pentaphyllum* with those of *C. lanata*\textsuperscript{87}, *C. pepo*\textsuperscript{88}, *C. sativus*\textsuperscript{89}, and *D. glomerata*\textsuperscript{90}, and calculated synonymous nucleotide substitutions (*Ks* values) between syntenic genes. The *Ks* values between syntenic homologous gene pairs were calculated using WGDI v0.5.1\textsuperscript{91} and the timings of WGD events were calculated with the following formula: divergence rate \( T = Ks/2r \textsuperscript{92}. \)

In addition, we assessed the contraction and expansion of gene families. First, species-specific gene families were filtered based on previous results for gene families. Second, CAFE\textsuperscript{93} was used to estimate gene family contraction and expansion in species based on gene birth and apoptotic models, and combined with the divergence times. Finally, genes that significantly expanded or contracted (\( p \leq 0.05 \)) were selected for KEGG pathway enrichment.

**Transcriptome Sequencing and Data Analyses**

Total RNA was extracted using the ethanol precipitation protocol and CTAPBIOZOL reagent according to the instructions provided with the manual, and quantified with a Nano Drop and Agilent 2100 Bioanalyzer (Thermo Fisher Scientific, MA, USA) by BGI-Shenzhen (Shenzhen, China). Stranded RNA libraries were constructed for each sample and sequenced on the MGISEQ-2000 platform to generate paired end 150 bp reads (PE150). After sequencing, the data were filtered using SOAPnuke.

Clean reads were then mapped onto the *G. pentaphyllum* genome using HISAT2\textsuperscript{94}, and the data obtained were assembled using StringTie v2.1.4\textsuperscript{95}. The expression value of each gene measured as the fragments per kilobase of transcript per million mapped fragments (FPKM)\textsuperscript{96} was calculated using the R package ballgown. The correlations were analyzed among 16 sample replicates (four young leaves, four stems, three tendrils, four flowers, and one fruit) using the R package Heatmap2. All of the assembled unigenes were annotated by aligning to the *G. pentaphyllum* genome. Genes involved in the gypenoside biosynthesis pathway were screened out and differential expression values were calculated using the limma package. Genes with adjusted \( \log_{2}FC \) \( > 1 \) and \( p \) value < 0.05 were defined as DEGs. TBtools\textsuperscript{97} was employed to create an expression-level heatmap for different tissues based on \( \log_{10}(\text{FPKM}) \) data.

**GBS Sequencing and SNP Calling**

The GBS library was constructed using a one-enzyme system with the restriction enzyme MseI for cutting and adjusted with the enzyme EcoRI, before sequencing on the HiSeq X10 platform using a 150 bp paired end protocol to obtain a large number of raw reads. The clean data were obtained after strict filtration by removing sequences with barcodes, adapters, low quality bases (\( \leq 5 \) bp), and unmeasured base N (10% of single reads) from the raw reads. The two technical replicates for each sample were merged before variant calling as described by Wong et al.\textsuperscript{98}.

According to the sequence similarity, the remaining high quality reads were mapped onto the *G. pentaphyllum* genome using BWA\textsuperscript{68} and the alignment results were sorted with Samtools\textsuperscript{69}. SNP calling from multiple samples was performed using Genome Analysis Toolkit (GATK) v3.4\textsuperscript{99}.  


We obtained SNPs by using a simplified approach that only retained biallelic SNPs, as described previously for *Medicago sativa*\(^2\). Loci with low polymorphism (MAF < 0.05) and that exhibited significant deviation from Hardy–Weinberg equilibrium \((p < 0.01)\) were filtered from the data set using Vcftools\(^1\) and Plink v1.7\(^1\). Finally, successfully genotyped loci in at least 95% of cases and present in all individuals were retained for further population genetics analyses.

**Population Genomics Analyses**

The population genetic diversity and structure were investigated with the method described in Note S2. An individual-based ML phylogenetic tree was built using RAxML v8.0.5\(^2\). We used the GTRGAMMA model with ascertainment bias correction and a rapid bootstrap procedure with 1000 replicates. PCA was performed with GCTA v1.26\(^3\) and the first three eigenvectors were plotted. AMOVA was conducted using Arlequin v3.5\(^4\) based on the identified SNPs and the data sets determined with the ADMIXTURE analysis results. Moreover, we used the PSMC\(^5\) method with whole genome data to infer the history of effective population size changes in *G. pentaphyllum* on older timescales. The SMC++\(^6\) method tailored to handle large data sets to determine demographic changes on recent timescales was also applied to reconstruct the recent demographic history using GBS data. We employed a universal per-generation mutation rate of 1.5e–8 mutations per base pair for all dicots\(^7\)-\(^9\). In addition, according to our field investigations over many years, a constant generation time of 5 years was used for demographic analysis.

**Niche Comparison Analyses Based on Environmental Space (E-space) and Geographical Space (G-space)**

In order to capture the different ecological spaces occupied by the three groups, we performed comparative analyses in both the environmental (E) and geographical (G) spaces, which complemented each other in niche comparison analyses\(^1\)-\(^1\). The historical and present geographical distribution ranges of each *G. pentaphyllum* group were recovered from population sampling sites using MAXENT v.3.3.3k\(^1\). Models were developed based on their present geographic distributions and current environmental factors, and then projected onto the LIG (130 ka) and the LGM (21 ka) periods. Six climatic variables identified in a previous study\(^1\) were selected as data predictors (Note S3), and bioclimatic variables were downloaded from the WorldClim website. The LGM data used in this study were under the Community Climate System Model\(^1\). In addition, kernel density plots for each group according to the six environmental variables were generated using the R package ggplot2.

To evaluate the degree of niche overlap, occurrence and climate data were translated into environmental axes (PCA-env) by PCA using R scripts\(^1\)-\(^1\). The ecological backgrounds of the three groups (Group N, Group SW, and Group SE defined by the ADMIXTURE program) were selected from a minimum convex polygon with a buffer size of 0.3 degrees\(^1\). To construct the available environmental space for the two principal axes, the values of the six climatic factors were extracted and the densities of population occurrences were calculated using a kernel smoothing method\(^1\). Finally, we obtained multiple range PCA-env plots to represent all available climates and occupied conditions for occurrence densities of 20% and 100%. Moreover, niche overlaps between groups were characterized by niche stability, niche unfilling, and niche expansion\(^1\)-\(^1\).
ACKNOWLEDGMENTS

This study was supported by the National Natural Science Foundation of China (Nos. 32000256, 31900273, 31770229, and 31270364), a Project funded by the China Postdoctoral Science Foundation (No. 2020M673627XB), and the Key Program of Research and Development of Shaanxi Province (No. 2022ZDLSF06-02).

DATA AVAILABILITY

The whole genome sequence data including NGS short reads, Nanopore long reads, Hi-C interaction reads, transcriptome data, and genome file have been deposited in the NCBI database under BioProject: PRJNA810200.

URLs

SOAPnuke v1.6.5 (https://github.com/BGI-flexlab/SOAPnuke)
Medaka v0.7.1 (https://github.com/nanoporetech/medaka)
Samtools v1.9 (https://github.com/samtools)
BamDeal v0.22 (https://github.com/BGI-shenzhen/BamDeal)
RepeatMasker (http://repeatmasker.org)
RepeatProteinMask (http://www.repeatmasker.org/cgi-bin/RepeatProteinMaskRequest)
RepeatScout v 1.0.6 (http://bix.ucsd.edu/repeatscout/)
RepeatModeler-2.1 (http://www.repeatmasker.org/RepeatModeler/)
KEGG (http://www.genome.jp/kegg/)
GO (http://www.geneontology.org/)
SwissProt (http://www.uniprot.org/)
InterPro (https://www.ebi.ac.uk/interpro/)
TrEMBL (http://www.uniprot.org/)
TreeBeST (https://github.com/Ensembl/treebest/)
TimeTree (http://www.timetree.org/)
WorldClim (http://www.worldclim.org)

CONFLICT OF INTEREST

No conflict of interest declared.

AUTHOR CONTRIBUTIONS

Zhao G. F. and Li Z. H. conceived and designed the study. Zhang X., Yang J., Zhang H., and Zhao Y. M. collected the samples. Zhang X. and Zhao Y. H. interpreted the assembly and annotations and contributed to the genome evolution analyses. Zhao Y. H. performed the comparative transcriptome analyses. Zhang X., Kou Y. X. and Zhao Z. conducted the population genomic analyses based on genotyping-by-sequencing data. Chen X. D. analyzed the niche comparisons data. Zhang X. wrote the draft manuscript and Li Z.H. then revised the English content of the manuscript. The final manuscript has been read and approved by all authors.

REFERENCES


24 Shen, C., et al. (2020). The chromosome-level genome sequence of the autotetraploid alfalfa and resequencing of core germplasms provide genomic resources for alfalfa research. Molecular Plant. 13 (9): 1250 - 1261.


39 Ma, L. L., et al. (2020). The genome and transcriptome analysis of snake gourd provide insights into its evolution and fruit development and ripening. Horticulture Research. 7: 199.


18

69 Durand, N. C., et al. (2016). Juicer Provides a One-Click System for Analyzing


88 Montero-Pau, J., José B., Bombarely, A., Ziaroso, P., & Joaquín C. (2018). De-novo assembly of zucchini genome reveals a whole genome duplication associated with the origin of the
cucurbita genus. Plant Biotechnology Journal. 16(6).

89 Wóycicki, R., et al. (2011). The genome sequence of the north-european cucumber (Cucumis sativus L.) unravels evolutionary adaptation mechanisms in plants. Plos One, 6 (7), e22728.


Table 1. Statistics and assembly BUSCO evaluation of the *G. pentaphyllum* genome.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate of genome size</td>
<td>608.95 Mb</td>
</tr>
<tr>
<td>GC content</td>
<td>32.95%</td>
</tr>
<tr>
<td>N50 length (contig)</td>
<td>5.05 Mb</td>
</tr>
<tr>
<td>Number of contigs (&gt;=2kb)</td>
<td>158</td>
</tr>
<tr>
<td>N50 length (scaffold)</td>
<td>56.38 Mb</td>
</tr>
<tr>
<td>Number of scaffolds (&gt;=2kb)</td>
<td>18</td>
</tr>
<tr>
<td>Tandem repeats</td>
<td>45,479,485 bp (7.47%)</td>
</tr>
<tr>
<td>RepeatMasker</td>
<td>70,556,049 bp (11.59%)</td>
</tr>
<tr>
<td>RepeatProteinMask</td>
<td>57,563,766 bp (9.45%)</td>
</tr>
<tr>
<td>De novo</td>
<td>350,543,161 bp (57.57%)</td>
</tr>
<tr>
<td>Total</td>
<td>366,739,319 bp (60.23%)</td>
</tr>
<tr>
<td>Number of genes</td>
<td>27,418</td>
</tr>
<tr>
<td>Average mRNA length</td>
<td>3,397 bp</td>
</tr>
<tr>
<td>Average CDS length</td>
<td>2,350 bp</td>
</tr>
<tr>
<td>Average exon per gene</td>
<td>40</td>
</tr>
<tr>
<td>Average exon length</td>
<td>233 bp</td>
</tr>
<tr>
<td>Average intron length</td>
<td>115 bp</td>
</tr>
<tr>
<td>Complete BUSCOs</td>
<td>1310 (95.3%)</td>
</tr>
<tr>
<td>Complete and single-copy BUSCOs</td>
<td>1284 (93.4%)</td>
</tr>
<tr>
<td>Complete and duplicated BUSCOs</td>
<td>26 (1.9%)</td>
</tr>
<tr>
<td>Fragmented BUSCOs</td>
<td>9 (0.7%)</td>
</tr>
<tr>
<td>Missing BUSCOs</td>
<td>56 (4%)</td>
</tr>
<tr>
<td>Total BUSCO groups searched</td>
<td>1,375 (100%)</td>
</tr>
</tbody>
</table>
Table 2. Genomic footprint of transposable elements in the genome of *G. pentaphyllum*.

<table>
<thead>
<tr>
<th>Repbase TEs</th>
<th>TE protiens</th>
<th>De novo</th>
<th>Combined TEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>Length (Bp)</td>
<td>% in genome</td>
<td>Length (Bp)</td>
</tr>
<tr>
<td>DNA</td>
<td>12740163</td>
<td>2.09</td>
<td>4422274</td>
</tr>
<tr>
<td>LINE</td>
<td>3445353</td>
<td>0.57</td>
<td>2186354</td>
</tr>
<tr>
<td>SINE</td>
<td>40999</td>
<td>0.01</td>
<td>0</td>
</tr>
<tr>
<td>LTR</td>
<td>54777642</td>
<td>9.00</td>
<td>50956807</td>
</tr>
<tr>
<td>Total</td>
<td>71004157</td>
<td>11.66</td>
<td>57565435</td>
</tr>
</tbody>
</table>

Table 3. Analysis of molecular variance (AMOVA) based on GBS data of *G. pentaphyllum*.

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>d.f.</th>
<th>Sum of squares</th>
<th>Variance components</th>
<th>Percentage of variation</th>
<th>Fixation Indices</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 Groups</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Among groups</td>
<td>1</td>
<td>639.47</td>
<td>7.15128 Va</td>
<td>5.44</td>
<td>FSC : 0.39012**</td>
</tr>
<tr>
<td>Among populations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>within groups</td>
<td>30</td>
<td>7070.902</td>
<td>48.45045 Vb</td>
<td>36.89</td>
<td>FST : 0.42333**</td>
</tr>
<tr>
<td>Within populations</td>
<td>75</td>
<td>5680.75</td>
<td>75.74333 Vc</td>
<td>57.67</td>
<td>FCT : 0.05445**</td>
</tr>
<tr>
<td>Total</td>
<td>106</td>
<td>13391.121</td>
<td>131.34506</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Groups</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Among groups</td>
<td>2</td>
<td>1068.761</td>
<td>8.16692 Va</td>
<td>6.26</td>
<td>FSC : 0.38077**</td>
</tr>
<tr>
<td>Among populations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>within groups</td>
<td>29</td>
<td>6641.61</td>
<td>46.57533 Vb</td>
<td>35.69</td>
<td>FST : 0.41953**</td>
</tr>
<tr>
<td>Within populations</td>
<td>75</td>
<td>5680.75</td>
<td>75.74333 Vc</td>
<td>58.05</td>
<td>FCT : 0.06259**</td>
</tr>
<tr>
<td>Total</td>
<td>106</td>
<td>13391.121</td>
<td>130.48559</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
FIGURE LEGENDS

Figure 1 Plant morphology, karyotypes, assembled genome information, and evolutionary analyses of *G. pentaphyllum*. (a) Wild plant and habitat of *G. pentaphyllum*. (b) Venn diagram for the genome annotated based on five databases. (c) Circos graph showing the genome characteristics from the outer circle to the inner ring (A–G): (A) circular representation of the 11 pseudochromosomes featured in 10-Mb intervals across the genome; (B–D) distributions of (B) Ty3-gypsy long terminal repeat retrotransposons (LTR-RTs), (C) Ty1-copia LTR-RTs, and (D) all transposable elements; (E) gene density, (F) GC distribution, and (G) intraspecific collinearity. (d) Flow cytometry analysis. (e) Anatomical sections of *G. pentaphyllum* chromosomes. (f) Hi-C chromatin interaction heatmap for 11 pseudochromosomes in the *G. pentaphyllum* genome. (g) Inferred phylogenetic tree, divergence time, and gene family expansions and contractions. The numbers at the node positions represent the divergence time of each species in millions of years. The numbers in parentheses indicate the confidence interval for the divergence time, which can be used to estimate the divergence time of target species and other species. The yellow points are the calibration times used to correct the time of species divergence. The green, red, and blue sections in the pie graph indicate the expanded, contracted, and other gene families, including extinctions and no change gene families, respectively. The red and green numbers represent the exact precise quantities of expanded and contracted gene families, respectively. (h) Clusters of orthologous and paralogous gene families in *G. pentaphyllum* and 11 other sequenced plant genomes. (i) Whole-genome duplication (WGD) analysis of *G. pentaphyllum* and genomes of related species.

Figure 2 Comparative transcriptome analysis of genes involved in the gypenoside biosynthetic pathway. The expression value of each identified candidate gene is colored in log_{10}(FPKM) in five tissues: fruit, flower, stem, leaf, and tendril. Low to high expression is indicated by the change in color from blue to red.

Figure 3 Genetic structure, demographic history, and global climatic space constructed over all background areas and realized niches of *G. pentaphyllum*. (a) Spatial genetic structure of *G. pentaphyllum* based on 32 populations within the natural species range in China. The pie chart colors indicate the probability of sample assignment based on SNPs analyzed using ADMIXTURE at the most likely K = 3. Standard maps produced by the Ministry of Natural Resources were used as the base maps with drawing approval numbers GS(2016)1554 and GS(2016)1606. (b) Results of ADMIXTURE analyses for 107 *G. pentaphyllum* individuals at K = 2 and K = 3. (c) Non-bootstrapped maximum-likelihood phylogenetic tree based on SNPs in 107 *G. pentaphyllum* individuals. Color coding of the branches reflects the structure of genetic groups at K = 3. (d) Past effective population size history of *G. pentaphyllum* assessed by PSMC. (e) Demographic changes on recent timescales established for *G. pentaphyllum* by using SMC++. (f) Plot of cross-validation errors for ADMIXTURE runs with K values ranging from 1 to 8. (g) Three-dimensional principal component analysis (PCA) results for *G. pentaphyllum*. (h) The 20% occurrence density includes the contribution and direction of each variable to the first two components of PCA-env. (i) An occurrence density of 100% is denoted by the thin line and 100% of the available climatic background by the thick line. (j–o) Kernel density plots for six environmental parameters for *G. pentaphyllum* showing niche differences among the three groups.