Multi-omics reveal differentiation and maintenance of dimorphic flowers in an alpine plant on the Qinghai-Tibet Plateau

Mingjia Zhu[1], Zhenyue Wang[1], Yongzhi Yang[1], Zefu Wang[2], Wenjie Mu[1], and Jianquan Liu[2]

[1]Lanzhou University
[2]Sichuan University

December 1, 2021

Abstract

? Dimorphic flowers growing on a single individual plant play a critical role in extreme adaption and reproductive assurance in plants and have high ecological and evolutionary significance. However, the omics bases underlying such a differentiation and maintenance remain largely unknown. We aimed to investigate this through genomic, transcriptome and metabolomic analyses of dimorphic flowers in an alpine biennial, Sinoswertia tetraptera (Gentianaceae). ? A high-quality chromosome-level genome sequence (903 Mb) was first assembled for S. tetraptera with 31,359 protein-coding genes annotated. Two rounds of recent independent whole-genome duplication (WGD) were revealed. More than 10% of the novel genes from the recent species-specific WGD were found to be differentially expressed in the two types of flowers, and this may have helped contribute to the origin of this innovative trait. ? Other contrasting gene expression between flowers included that related to flower development and color, hormones, and iridoid biosynthesis. Metabolomic analyses similarly suggested differential concentrations of both hormones and iridoids in the two types of flowers. The interactions between multiple genes may together lead to contrasting morphology and open versus closed pollination of the dimorphic flowers in this species. ? A total of 56 candidate genes were identified from the known iridoid biosynthesis-related pathways. Two hub genes were found to play an essential role in transferring intermediate products between leaves and flowers during iridoid biosynthesis.

Introduction

Polyphenism is a unique type of phenotypic plasticity, in which the outputs are not continuous but relatively discrete, arising from the same genotype (Mayr, 1963; Moran, 1992; C.-H. Yang & Andrew Pospisilik, 2019). Its diverse traits and importance in conferring ecological fitness have been widely acknowledged (Darwin, 1897; Simpson, Sword, & Lo, 2011) in both animals and plants (Abouheif & Wray, 2002; Fawcett et al., 2018; Yiyang Liu et al., 2021; Zhang et al., 2021). As one of the most widespread polyphenisms (Joly & Schoen, 2021), dimorphic flowers evolved independently in roughly 700 species from 50 families of plants (Culley & Klooster, 2007). Such flowers are usually chasmogamous (CH) and cleistogamous (CL) growing on a single plant and having contrasting shapes, colors, and smells (Campbell, Quinn, Cheplick, & Bell, 1983; Lord, 1981). CH flowers (hereafter CHs) have bright and colorful petals and nectaries and remain open for cross-pollination, while CL flowers (hereafter CLs) have green petals and remain closed for self-fertilization (Darwin, 1897; Lord, 1981). In addition, CLs are always smaller and have a simpler structure than CHs, leading to lower costs and an automatic transmission advantage (Wang, Du, & Wang, 2017). Therefore, as a 'pessimistic strategy', CLs can ensure reproductive success under harsh or uncertain conditions (Schnee & Waller, 1986; Waller, 1980). However, selfing through CLs rapidly leads to inbreeding depression and adversely affects genetic load (building up harmful mutations) (Ansaldi, Weber, & Franks,

2018; Charlesworth & Charlesworth, 1987). Thus, outcrossing CHs can effectively increase recombination and overcome these weakness (Culley & Klooster, 2007; Culley & Wolfe, 2001). Such a trade-off with a mixed mating system provides high reproductive assurance, allowing plants to survive unpredictable and extreme environments (Ansaldi et al., 2018; Koontz, Weekley, Haller Crate, & Menges, 2017).

In addition to ecological significance, it would be interesting to know what genes and their expressions have led to the differentiation and maintenance of such dimorphic flowers with the same genotype (Ansaldi et al., 2018; Morinaga et al., 2008). Two recent studies have investigated this genetic differentiation through sequencing the genome and examining gene expression of the dimorphic flowers (Yiyang Liu et al., 2021; Zhang et al., 2021). For *Amphicarpaea edgeworthii* (Fabaceae) with dimorphic flowers, the identified genes with contrasting expressions between aerial CH and subterranean CL flowers were mainly related to MADS-box genes (Yiyang Liu et al., 2021). Research on *Cleistogenes songorica* (Gramineae) with dimorphic flowers suggests miRNA, *MYB*transcription factors, and targeted genes are involved in the differential development of the highly reduced CH and CL flowers in this grass (Zhang et al., 2021). However, the typical structures of the dimorphic flowers of these two species differ from most species with aerial dimorphic flowers (Campbell *et al.* , 1983; Culley and Klooster, 2007). In this study, we used multi-omics data to examine differentiation and maintenance of dimorphic flowers in a more typical species, *Sinoswertia tetraptera* (Gentianaceae), an endangered, alpine biennial restricted to the Qinghai–Tibet Plateau (L. Yang, Zhou, & Chen, 2011). In the entire family Gentianaceae only this monotypic genus has a mixed mating system with both CH and CL flowers on the top or basal stem of a single plant (Figure 1a) and CH flowers may disappear in some plants in the high-altitude extremes (T. He, Liu, & Liu, 2013). The open CH flowers are pale-blue and large with distinct nectaries, while the closed CLs are green without nectaries. Such contrasting shapes and colors are similar to most dimorphic flowers of other species (Culley & Klooster, 2007). In addition, this species has been used as a traditional Tibetan medicine since the 6^{th} century BCE (Rao *et al.* , 2010) and is rich in iridoid compounds (Brahmachari et al., 2004; Yue Liu et al., 2017). It remains unknown whether the two types of flowers contain the same or different concentrations of iridoid compounds and related gene expressions.

We assembled a chromosome-level *de novo* genome of *S. tetraptera* using long Nanopore reads, Illumina short reads, and Hi-C data. By comparative genomic analyses, we first explored the genome evolution of *S. tetraptera* , the first representative of the family Gentianaceae. Then, based on transcriptome and metabolome analyses, we further examined the omics differentiation and maintenance of dimorphic flowers in *S. tetraptera* . Finally, we identified candidate genes involved in iridoid biosynthesis according to the weighted gene co-expression network analysis (WGCNA) and gene expressions. Our data provide essential insights into how CH and CL flowers are differentially maintained and iridoids are synthesized in this alpine plant.

2. Materials and methods

2.1 Plant materials and sequencing

For genome sequencing, one adult plant of*S. tetraptera* was collected from Xining, Qinghai Province, China (N37°15'8, E101deg22'18). Fresh leaves were collected and snap-frozen in liquid nitrogen for DNA isolation. The modified CTAB method (Doyle & Doyle, 1987) was used to extract the genomic DNA. A Nanopore library was constructed following the Nanopore library construction protocol. A total of 152.6 Gb long reads were generated using the PromethION sequence platform (Oxford Nanopore Technologies, [ONT]) (Table S1). A paired-end library with 350 bp insert fragments was constructed and then 51.8 Gb of raw data were produced using the Illumina HiSeq 2000 platform (Table S1). We also built a Hi-C library from young leaves as described previously (Y. Yang et al., 2020) and obtained 121.31 Gb Hi-C reads using the Illumina HiSeq 2000 platform (Table S1).

In addition, seven tissues (roots, stems, cCH [closed chasmogamous] flowers, bCH [blooming chasmogamous]

flowers, CL [cleistogamous] flowers, and leaves from the branch of CH [termed 'CH leaf'] and CL [CL leaf] flowers, respectively) from the same plant were collected and immediately frozen in liquid nitrogen. For each tissue, three samples were collected as biological replicates for RNA-sequencing (RNA-seq) and metabolome analysis. The total RNA was extracted using a QIAGEN RNeasy plant mini kit for each sample. The RNA-seq libraries were then constructed with a TruSeq RNA library preparation kit (Illumina). A total of 96.24Gb RNA-seq reads for all 21 libraries were obtained from the HiSeq 2000 platform (Table S1).

2.2 Genome size assessment

The clean reads (152 Gb) from the 350 bp insert-size Illumina library were used to estimate the genome size based on the 17-mer method. Jellyfish (Marcais & Kingsford, 2011) was used to generate the k -mer frequency distribution. Genomescope (Ranallo-Benavidez, Jaron, & Schatz, 2020) was employed to estimate the genome size. The k -mer number was 33,414,462,943, with a peak depth of 34. The estimated genome size of *S. tetraptera* was ~ 982.78 Mb (Figure S1, Table S1).

2.3 Genome assembly and quality control

We used NextDenovo v2.4.0 (https://github.com/Nextomics/NextDenovo) to *de novo* assemble the genome with ONT long reads (100x). First, the NextCorrect module was applied to correct the raw reads, then the preliminary genome assembly was generated by the NextGraph module. Purge Haplotigs (Roach et al., 2018) were used to identify and remove the candidate duplicate haplotypes to manually curate the heterozygous assemblies. Racon (Vaser, Sović, Nagarajan, & Šikić, 2017) v1.4.20 was then employed to polish the assembly for two rounds with the corrected ONT long reads (Figure S2 and S3). Finally, we used Nextpolish (J. Hu, Fan, Sun, & Liu, 2020) v1.3.1 for two rounds of assembly polishing based on Illumina short reads (100×) and then we generated the final genome assembly.

We anchored the genome assembly to the chromosome level using the Hi-C data. HiC-Pro (Servant et al., 2015) was employed to control the raw data with default parameters. Bowtie2 (Langmead & Salzberg, 2012) was used to map the Hi-C reads to the assembled genome. The unique mapped reads were extracted, with duplicates excluded, by HiC-Pro. Finally, we used LACHESIS (Burton et al., 2013) to cluster, reorder, and orientate the corrected contigs onto pseudo-chromosomes based on the interaction level.

To assess the quality of our assembly, whole-genome sequencing (NGS) reads and assembled transcripts were mapped to the genome by BWA (H. Li, 2013) v0.7.17 and HISAT2 (D. Kim, Langmead, & Salzberg, 2015) v2.1.0, respectively. Benchmarking Universal Single-Copy Orthologs (BUSCO) (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015) was also employed to assess the completeness of the assembly based on the dataset of embryophyta_odb10.

2.4 Genome annotation

Repetitive elements were predicted in the genome of *S. tetraptera* . We used TRF (Benson, 1999) and MISA (Thiel, Michalek, Varshney, & Graner, 2003) to identify the tandem repeats and simple sequence repeats (SSRs), respectively. Transposable elements (TEs) were then identified based on *de novo* and homology-based strategies. RepeatMasker (Tarailo-Graovac & Chen, 2009) v4.0.7 was used to run a homology search for known repeat sequences against the Repbase database v22.11 (Jurka et al., 2005). RepeatModeler (Jurka et al., 2005) v2.0.10 was employed to predict the TEs based on the *de novo* method. Finally, all identified repetitive elements were merged for subsequent analyses.

Protein-coding genes were then predicted in the repeat-masked *S. tetraptera* genome based on integrated strategies. The RNA-seq reads derived from the seven tissues (TableS2) were assembled using Trinity v2.6.6

(Grabherr et al., 2011) in the *de novo* -based and genome-guided modes, respectively. For transcriptome-based prediction, the assembled transcripts produced in the two different ways were combined and further aligned to the genome by PASA v2.1.0 to obtain the gene structures. For homology-based prediction, protein sequences of seven species (*Arabidopsis thaliana* (Kaul et al., 2000), *Vitis vinifera* (Jaillon et al., 2007), *Solanum melongen*(Barchi et al., 2021), *Calotropis gigantea* (Hoopes et al., 2018), *Coffea canephora* (Denoeud et al., 2014), *Catharanthus roseus* (Kellner, Kim, Clavijo, Hamilton, Childs, Vaillancourt, Cepela, Habermann, Steuernagel, Clissold, McLay, et al., 2015) and *Oryza sativa* (J. Yu et al., 2002)) were selected and aligned against the genome of *S. tetraptera* using GeMoMa (Keilwagen, Hartung, & Grau, 2019) v1.6.1. Augustus (Stanke, Steinkamp, Waack, & Morgenstern, 2004) v3.3.3, GlilmmerHMM (Majoros, Pertea, & Salzberg, 2004), and GeneScan (Burge & Karlin, 1997) were then employed for the *ab initio* gene prediction. The assembled transcripts of *S. tetraptera* were used as the training set for Augustus. Finally, all forecasts produced by different strategies were integrated into a final gene set using EVidenceModeler v1.1.1(EVM) (Haas et al., 2008). BUSCO was used to assess the completeness of gene prediction.

For function annotation of the predicted protein-coding genes, three public databases – Swiss-Port, TrEMBL (Boeckmann et al., 2003), and NR (Coordinators, 2016) – were used to search against BLAST (Rédei, 2008). Then we used InterProScan (Quevillon et al., 2005) to predict information relating to protein domains. The Gene Ontology (GO) terms were retrieved by the pipeline of Blast2GO v2.5 (Conesa et al., 2005). The pathway information for each gene was assigned by the KEGG database (Conesa et al., 2005).

2.5 Phylogenetic analysis and divergence time estimation

We used OrthoMCL (L. Li, Stoeckert, & Roos, 2003) to identify the gene families (orthologous and paralogous groups) in *S. tetraptera* and the other eleven species: *O. sativa* , *C. gigantea, V. vinifera, C. canephora, C. roseus, Gelsemium sempervirens* (Franke et al., 2019), *Gardenia jasminoides* (Xu et al., 2020), *Eucommia ulmoides* (Wuyun et al., 2018), *Capsicum annuum* ,*Aquilegia coerulea* (Filiault et al., 2018), and*Camellia sinensis* (Xia et al., 2020). A total of 485 single-copy gene groups were identified and extracted. For each gene, the protein sequences were aligned by MAFFT v7.467 (Katoh & Standley, 2013), and then the coding sequences (CDS sequences) were aligned by PAL2NAL v.14 (Suyama, Torrents, & Bork, 2006) under the guidance of corresponding protein alignments. For all CDS alignments, the conserved sites were extracted to generate the concatenated sequences for each species. Finally, the phylogenetic tree was constructed by IQ-TREE (Nguyen, Schmidt, Von Haeseler, & Minh, 2015) v1.6.9, with the best-fitted substitution model produced by ModelFinder (Kalyaanamoorthy, Minh, Wong, Von Haeseler, & Jermiin, 2017) and 1,000 replicates (-bb 1000 -m MFP).

MCMCTREE in the PAML v4.9 package (Z. Yang, 2007) was employed to date the divergence times. Two fossil constraints and a soft-bound maximum were used at the split node of (1) monocots-eudicots (130-190 million years ago [Ma]) (H. T. Li et al., 2019); (2) asterids-rosids (116–126 Ma) (H. T. Li et al., 2019); and (3) *C. annuum* - *C. canephora* (85-91 Ma) (Hedges, Marin, Suleski, Paymer, & Kumar, 2015). Finally, we used CAFÉ (De Bie, Cristianini, Demuth, & Hahn, 2006) to explore the expanded and contracted gene families.

2.6 Whole-genome duplication (WGD) analysis

WGD analysis was performed using four genomes: *V. vinifera*(eudicots; Vvi), *C. canephora* (asterids; Cca), *G. jasminoides* (asterids; Gja), and *S. tetraptera* (asterids; Ste). Synteny analyses between and within species were surveyed using wgdi (Sun et al., 2021). Rectangles with different colors highlighted the collinear blocks containing at least 10 gene pairs. For each colinear gene pair, synonymous substitutions per synonymous site (*Ks*value) were calculated using the Nei-Gojobori (NG) approach implemented in PAML (Z. Yang, 2007) v4.9. The median *Ks* values for each collinear block were extracted to estimate the Ks distributions further

using wgdi. The collinear blocks were classified into different groups according to their *Ks* values, then the genes produced were identified by the two most recent WGDs using wgdi (with "-a" parameter).

2.7 Gene family analyses

We identified the homologous gene families involved in flowering time, flower development, and flavonoid and carotenoid biosynthesis in*S. tetraptera* . The known genes from each family were downloaded as the query to search against the *S. tetraptera* genome using BLASTP (Rédei, 2008). HMMER (Eddy, 2011) was then used to search for previously known domains from corresponding gene families for the candidate sequences. The candidate genes not harboring the domains searched for were removed. All the query sequences and the previously known domains are summarized in Table S23-24 and Table S27-28. For each gene family, MAFFT was used to align the protein sequences. IQ-TREE was used to construct the phylogenetic trees with default parameters (Nguyen et al., 2015), and further illustrated by EVOLVIEW (Z. He et al., 2016). We also predicted the transcription factors in the *S. tetraptera*genome using PlantRegMap (Tian, Yang, Meng, Jin, & Gao, 2020) and the PlantTFDB database (Jin et al., 2017). In addition, clusterProfiler v3.6.0 (R package) (G. Yu, Wang, Han, & He, 2012) was used to analyze the enrichment of gene families in this study.

2.9 Gene expression and weighted gene co-expression network analysis

A total of 21 transcriptomes from seven tissues (three biological replicates for each tissue) were used to obtain the gene expression and to perform the weighted gene co-expression network analysis (WGCNA). For each sample, RNA-seq short reads were filtered using fastp (Chen, Zhou, Chen, & Gu, 2018) with default parameters, and mapped to the *S. tetraptera* genome by HISAT2 (D. Kim et al., 2015). The transcripts per million (TPM) values for each gene were then extracted to measure their expression level by StringTie (Pertea et al., 2015). Differentially expressed genes (DEGs) between different tissues were identified by DESeq2 (R package) (Love, Huber, & Anders, 2014). The candidate genes with at least two-fold differential expression levels in various tissues and an FDR cut-off value of 0.05 were identified as DEGs. The weighted gene co-expression network analysis was then performed by WCGNA (R package) (Langfelder & Horvath, 2008). The generated network was visualized using Cytoscape (Smoot, Ono, Ruscheinski, Wang, & Ideker, 2011) v3.7.2.

2.10 Metabolomics analysis

The samples from the seven tissues were harvested as previously described. For each sample, 20 mg of powder was prepared and further extracted with 400 μL of 80% aqueous methanol at 4 , followed by centrifugation for 10 min at 12,000 rpm. LC-MS analysis was performed using the Waters Acquity UPLC System connected to an AB SCIEX 5500 QQQ-MS.

Gradient elution was achieved on a Waters Acquity UPLC BEH C18 column (100mm*2.1mm, 1.7μm) with water containing 0.1% formic acid (solvent A) and acetonitrile (solvent B) at a flow rate of 0.30 mL/min (X. Yu et al., 2020). The column temperature was maintained at 40. The gradient elution program was as follows: 1-10%B (0-1min),11-60%B (2-5min), 60-90%B (5-7min), held at 99 %B (7-9min), and allowed to equilibrate for a further 3min before the next injection and the last 8min of the chromatogram solutions were discarded. The injection volume was 4μL. MS data were recorded with the following parameters: Ion source, ESI; IonSource temperature, 450; IonSource Gas1, 55arb; IonSource Gas2, 55arb; IonSpray voltage, 4500V; Curtain Gas, 35arb; Collision GAS, 7arb.

Components eluting from the UPLC-QQQ-MS system were processed in MultiQuant for data preprocessing with default settings, except that each sample was normalized to the internal standard (X. Yu et al., 2020). After filtering for outliers, the data were used for the subsequent statistical analysis.

3. Results and discussion

3.1 Genome assembly and annotation

A total of 22 Gb of Illumina short reads (100×) and 73 Gb of Oxford Nanopore Technologies (ONT) long reads (100 ×) were generated for*S. tetraptera* (Table S2). The contig-level assembly of*S. tetraptera* was 943 Mb in length (covering 96.03% of the estimated size), with 199 contigs and a contig N50 length of 4.9 Mb (Tables S3). Using121Gb (100×) of Hi-C data, we further anchored 95.76% of the assembly (903 Mb) onto six pseudochromosomes (Figure 1b, Table S4, Figure S4). The accuracy and completeness of the genome assembly were assessed according to the following: (1) 98.90% of NGS reads could be mapped to the assembly (Table S5); (2) 94%-98% of assembled transcripts could be mapped for more than 50% of the length (Table S6); (3) 96.50% (1326 out of 1375) Benchmarking Universal Single-Copy Orthologs (BUSCO) were fully present in the assembly (Table S7). These results indicated that the assembly of*S. tetraptera* was reliable with high completeness, continuity, and accuracy.

Around 70.88% of the *S. tetraptera* genome was identified as repetitive sequences, consisting of 69.55% interspersed repeats and 1.33% tandem repeats (Table S8). Long terminal repeats (LTRs) occupied the greatest proportion (47.51%), including 35.33% that were *Gypsy* elements and 11.70%*Copia* elements (Table S8). In addition, a total of 31,359 protein-coding genes were predicted in the genome, with an average gene length of 3,297 bp, an average exon sequence length of 224 bp, average exon number of 5.5 per gene, average intron length of 458 bp, and a GC content similar to the other previously reported Gentianales genomes (Figure S5-6, Table S9). Among all predicted protein-coding genes, 96.29% were functionally annotated by at least one database – SwissPort, TrEMBL, InterPro, GO, KEGG, Enogg-Mapper or NR (Table S10).

3.2 Phylogenetic analyses and evolution of gene families

To explore the phylogenetic relationship of *S. tetraptera* , we first performed the phylogenetic analyses with the other 11 species (Table S11). A total of 345,243 genes from the 12 species were assigned into 28,703 gene families (Figure S7,Table S12). For *S. tetraptera* , 25,573 genes could be clustered into 12,826 identified gene families (Table S14). A total of 485 groups of single-copy gene families were identified and used to construct the phylogenetic topology for all species. The phylogenetic positions of all species agreed well with previous studies (Figure 2a) (Chase et al., 2016; Y. Yang et al., 2020). A sister relationship between*S. tetraptera* and *G. sempervirens* was indicated by our results (Figure 2a). Their divergence time was dated to around 61.50 Ma (Figure 2a).

In addition, 3,555 gene families were found to expand in *S. tetraptera* . GO enrichment analysis revealed that these expanded gene families were significantly enriched ($P<0.01$, $Q<0.05$) in secondary metabolic processes, mainly involved in terpenoid and flavonoid biosynthesis (Figure 2b, Figure S8-9, Table S13-S16). This may contribute to the high content of Sinoswertiamatin, sweroside, gentiopicroside, and loganin in *S. tetraptera* , which may be relevant to the species' use in traditional Chinese medicine (Organization, 2002; Rao et al., 2010).

3.3 Whole-genome duplication (WGD)

We obtained synonymous substitution per site (*Ks*) distributions for each species based on the identified syntenic paralogues. Three rounds of WGD events were detected in *S. tetraptera,* different from the other previously reported Gentianales species (*C. canephora, G. sempervirens* , and *C. gigantea*), which all had only one WGD event (Figure 2c, Figure S10, Table S17-18). For all WGD events in *S. tetraptera* , the most ancient one (around 121-136 Ma) was shared by all species used in our study (three Gentianales species and *V. vinifera*), indicating the joint γ event shared by all core eudicots, as inferred in previous reports. The other two recent WGD events were species-specific in *S. tetraptera* and dated to 41-46 Ma and 67-75 Ma, respectively, the latter of which occurred shortly after the divergence between *S. tetraptera* and *G. sempervirens* (61.50Ma) (Figure 2c).

To further illustrate the WGD events in *S. tetraptera* , we undertook a collinearity analysis between *V. vinifera* and*S. tetraptera* . Based on the dot plots of paralogues, the observed syntenic depth ratios of 12:3 in the *S. tetraptera* - *V. vinifera* comparison indicated the occurrence of the joint γ event and two unique recent WGD events in*S. tetraptera* (Figure 2d). The same results were also revealed from the comparisons of *S. tetraptera* - *C. canephora* ,*S. tetraptera* - *G. jasminoides* , *S. tetraptera* -*G. sempervirens* , and *S. tetraptera* - *C. gigantea* , respectively (Figures S11-17). In addition, we identified a total of 2,269 genes that strictly originated from the most recent WGD event. We therefore examined their differentiated expressions between the two types of flowers, and found that more than 10% of these genes did exhibit contrasting expressions. They were functionally enriched in heterochronic development and growth, which may contribute to the development of the innovative dimorphic flowers in *S. tetraptera*(Figure 2b, Figure S18 and S19, Table S19-22).

3.4 Transcriptomic and metabolomic differentiation between CH and CL flowers

In addition to open and closed pollination of these two types of flowers, CH and CL flowers also have contrasting shapes, colors, and nectaries (L. Yang et al., 2011). The MADS-box gene family plays an essential role in floral organ development in all angiosperms (Ng & Yanofsky, 2001). We first examined their expressions in both dimorphic flowers. A total of 59 MADS-box genes were identified in *S. tetraptera* . They could further be classified into 12 clades, covering all of the identified clades in*O. sativa* and *A. thaliana* (Figure 3a, Figure S20) (Arora et al., 2007; Parenicova et al., 2003). Compared to *O. sativa* and *A. thaliana* , a conserved copy number of A-, B-, C- and E-class genes were detected in*S. tetraptera* , including six copies of *AP1* s (class A), two copies of *AP3* s and one copy of *PI* (class B), three copies of *AG* s (class C), and one copy of *SEP1* and *SEP3*(class E) (Figure 3a). Their conserved copy numbers indicated that they may play a key role in floral organ development (ES & EM, 1991; van Tunen, Eikelboom, & Angenent, 1993). However, the genes of the*AGL15/18* clade were highly expanded in *S. tetraptera*(Figure 3a) and a total of 21 more copies were identified from the tandem duplication.

Gibberellin (GA), Jasmonate acid (JA), and auxin (IAA) have been reported to play essential roles in regulating flower development (Ishiguro, Kawai-Oda, Ueda, Nishida, & Okada, 2001; Jibran, Tahir, Cooney, Hunter, & Dijkwel, 2017; Nagpal et al., 2005; Teotia & Tang, 2015). Our metabolomics analysis revealed the content of each differed significantly between CH and CL flowers (Figure 3b, Table S23). The contents of GA and JA in CHs were both significantly higher than those in CLs. We further assessed the expression level of genes related to GA, JA, and IAA regulated pathways, many of which belong to the MADS-box gene family (Table S24). *AGAMOUS* (*AG*) (E class of MADS-box gene family) can bind to the promoter of *DEFECTIVE IN ANTHER DEHISCENCE 1* (*DAD1*) and further positively regulate the content of JA (Ito *et al.* , 2007; Hu *et al.* , 2017). Both *AG*and *DAD1* were highly expressed in CH flowers, probably contributing to their significantly higher JA content (Figure 3c). JA is probably involved in flowering by regulating *AGL15/18* genes (MADS-box gene family) (Ishiguro et al., 2001; Jibran et al., 2017).

AGL15/18 genes have been reported to promote the expression of Gibberellin 2-oxidase 6 (*GA2ox6*) (Zheng, Zheng, Ji, Burnie, & Perry, 2016) and directly reduce the level of bioactive GAs by catalyzing their immediate precursors or inactive forms (Y.-X. Hu et al., 2017). The overexpression of *AGL15* can delay blooming in*A. thaliana* (Adamczyk, Lehti-Shiu, & Fernandez, 2007). We found*AGL15* and *GA2ox6* had lower expressions in CH flowers (Figure 3c), in which a substantially higher concentration of GA accumulated than in CL flowers (Figure 3b). GA may induce flowering by up-regulating *SQUAMOSA promoter binding protein-like3*(*SPL3*), *SPL4* , and *SPL5* genes and further promote flowering by targeting with *FRUITFUL* (*FUL*), *LEAFY (LFY)* , and *APETALA1* (*AP1*). *FUL* (A class of MADS-box gene family) can also promote the expression of *SPL4* to control flower formation (Torti et al., 2012). These genes (*SPL3* ,*SPL4* , *SPL5* , *AP1* , *LFY* , and *FUL*) were highly expressed in CH flowers (Figure 3c). In addition, the dosage of*AUXIN RESPONSE FACTOR6* (*ARF6*) and *ARF8* could quantitatively affect the timing of flower maturation by regulating JA accounts (Nagpal et al., 2005). They are inhibited by *AGL15/18*and increased IAA (Yang et al., 2006; Zheng et al., 2016), while IAA could also delay flowering (Ke et al., 2018; Lu et al., 2018). The expression of *ARF6/8* is also consistent with the IAA content in CL flowers (Figure 3b and 3c), which suggests that IAA may play a role in flower blooming through *ARF6/8* .

Moreover, CH flowers are significantly larger than CLs (P <0.01) (Figure 4a, Table S25). Cytokinin (CTK) may regulate floral organ size by catalyzing itself with *cytokinin oxidase/dehydrogenase 3* (*CKX3*) and *CKX5* (Bartrina, Otto, Strnad, Werner, & Schmülling, 2011). We found both CTK content and the expression level of *CKX3* and *CKX5* were distinctly different between CH and CL flowers. A significantly higher CTK content was detected in CH flowers than in CLs (Figure 4a). The expression levels of *CKX3* and *CKX5* in CH flowers were lower than those in CLs (Figure 4b). These differences may contribute to the contrasting sizes of CH and CL flowers.

In addition, the petals of CH flowers are more colorful and brighter than CLs to attract pollinators to *S. tetraptera* (He et al., 2013). The gene expression levels involved in biosynthesized carotenoids, anthocyanin, and flavonoids generally engaged in petal coloring were obviously different between the two dimorphic flowers, with high expression in CHs (Figure 4c and Table S26). Furthermore, petals of CH flowers have nectaries (Figure 4d), which secrete nectar to attract insects for pollination (He et al., 2013). We identified the*sugars-will-eventually-be-exported-transporters* (*SWEET*) gene family in the *S. tetraptera*genome (Figure S21), and it has been suggested that the *SWEET9* homolog is a significant sugar efflux transporter in plants (Lin et al., 2014). Most *SWEET* genes (including *SWEET9*) in *S. tetraptera* showed a higher expression level in CH than CL flowers (Figure 4d). However, three*SWEET* genes were highly expressed in CLs (Figure 4d). In fact, two of them, *SWEET17* and *SWEET2,* exhibited high expression levels in all tissues (Figure 4d; Figure S22), indicating that these two genes may be involved in the sugar efflux transporters in the whole plant (Guo et al., 2014; Klemens et al., 2013). The third gene, researched on *SWEET8* , was highly expressed in pistil donor ligand (S.-Y. Kim, Yu, Hong, Woo, & Ahn, 2013), which may contribute to the potential pistil differences between CH and CL flowers in *S. tetraptera* (Figure 4b; Figure S22). Previous studies have reported that*block of cell proliferation 1* (*BOP1*), *BOP2,* and*CRABS CLAW* (*CRC*) are involved in nectary development (Kram & Carter, 2009). All of these genes were highly expressed in CHs, suggesting their essential roles in the nectary development of these flowers (Figure 4d, Table S24).

3.5 Iridoids biosynthesis

Iridoids of *S. tetraptera* comprise loganin, Sinoswertiamatin, sweroside, and gentiopicroside. We measured their concentrations in seven tissues. The content of each iridoid in the seven tissues showed similar variation trends to those reported previously (Yue Liu et al., 2017). Four iridoids had the highest content in CL flowers out of the seven tissues examined, and their contents in these flowers were all obviously higher than those in leaves from the branch on which the CL flowers were growing (termed the "CL leaf") (Figure 5a, Table S27). The regulatory pathway of the biosynthesis of iridoids has been reported to consist of the downstream seco-iridoids pathway and the upstream 2-C-methyl-D-erythritol 4- phosphate (MEP) and mevalonate (MVA)

pathways (Kellner, Kim, Clavijo, Hamilton, Childs, Vaillancourt, Cepela, Habermann, Steuernagel, Clissold, Mclay, et al., 2015; Yue Liu et al., 2017; Vranová, Coman, & Gruissem, 2013). We identified a total of 56 candidate genes for these two biosynthesis pathways and assessed their expression levels in different tissues for subsequent analyses (Table S28-29). As the primary supplier of intermediate products in the biosynthesis of iridoids, the MEP pathway acts mainly in leaves (Oudin, Courtois, Rideau, & Clastre, 2007; Vranová et al., 2013). However, these iridoid products are always concentrated in the flowers of*Swertia mussotti* (Yue Liu et al., 2017) and *S. tetraptera* . Therefore, the intermediate products of these iridoids may be transported from leaves to flowers in *S. tetraptera* .

We reconstructed a weighted gene co-expression network for iridoid biosynthesis pathways based on intersections of DEGs containing 8067 genes (618 TFs and 7449 structural genes) for the seven tissues. A total of 12 modules were clustered, and module 8 (blue) was indeed related to leaves, while module 9 (turquoise) was linked to flowers (Figure 5b, Figure S23-24). Many of the previously identified 56 candidate genes for biosynthesis of iridoids were clustered into these two modules. The leaf-related module contained most genes belonging to the MEP pathway (Figure 5b). Most of them showed higher expression levels in leaves than in flowers (Figure 5c, Figure S25-26). However, the candidate genes in the MVA pathway were mainly clustered in the flower-related module and highly expressed in flowers (Figure 5b,5c, Figure S25-26). Similar patterns were also revealed from the correlation analysis between co-expression network modules and the measured iridoid contents (Figure S27). These different clustered networks and differentially expressed genes in the two tissues may suggest the potential genetic basis for the distinct function between leaves and flowers for iridoid production and transport. Furthermore, we found that the genes from the seco-iridoids pathway were distributed in both modules. Of these, two *SLS* genes were clustered as the hub genes (Figure 5b). They may play an essential role in transporting intermediate products during iridoid biosynthesis between different tissues in *S. tetraptera* .

4. Conclusion

In this study, we have reported the genome sequence for *S. tetraptera* . Based on this reference genome, we examined transcriptome differentiation of dimorphic flowers. In addition to MADS-box genes (Yiyang Liu et al., 2021), we revealed more distinctly expressed genes related to open versus closed pollination, nectary development, petal color, and bioactive compounds when comparing CH and CL flowers in*S. tetraptera* . It should be noted that we first found that the new genes derived from the species-specific WGD may have been involved in the evolution of such an innovative trait. In addition, we found contrasting concentrations of hormones and iridoids and the differential expression of related genes when comparing the two flower types. Therefore, the evolution and development of the aerial dimorphic flowers from numerous unrelated families (Campbell *et al.* , 1983; Culley and Klooster, 2007) may involve multiple but different genes despite the common ecological role of reproductive assurance in extreme habitats (Koontz *et al.* , 2017; Ansaldi *et al.* , 2018). Further genomic studies and comparisons of more species with dimorphic flowers are needed to examine how this innovative trait originated repeatedly in unrelated angiosperms. In addition, we identified candidate genes for iridoid biosynthesis in *S. tetrapterae* . Our co-expression analyses revealed two hub genes, which may be essential in transferring intermediate products during iridoid biosynthesis between leaves and flowers. This information may be very useful for artificially creating iridoids in cultivated crops in the future.

Acknowledgments

This work was supported equally by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB31000000), the Second Tibetan Plateau Scientific Expedition and Research (STEP) program (2019QZKK0502), the National Natural Science Foundation of China (grant numbers 31590821 and 91731301), National Key Research and Development Program of China (2017YFC0505203), and, the Fundamental Research Funds for the Central Universities (Grant No. lzujbky- 2019), and International Col-

laboration 111 Programme (BP0719040). We received support for computational work from the Big Data Computing Platform for Western Ecological Environment and Regional Development and Supercomputing Center of Lanzhou University.

Author contributions

J.L. designed this research. J.L. and M.Z. collected samples. M.Z., Z.W., W.M. and Y.Y. analyzed the data. M.Z. and J.L. wrote the manuscript with inputs from all authors. Z.W. and Y.Y. helped with revision of the manuscript.

Data availability

DNA sequencing data in our present study are available in BioProject of CNSA (https://db.cngb.org/cnsa/) under accession number CNP0002339. The RNA-seq data are available under accession number CNP0002339. In addition, the final genome assembly is public under accession number CNA0036281

References

Abouheif, E., & Wray, G. A. (2002). Evolution of the gene network underlying wing polyphenism in ants. *SCIENCE* , *297* (5579), 249–252. doi: 10.1126/science.1071468

Adamczyk, B. J., Lehti-Shiu, M. D., & Fernandez, D. E. (2007). The MADS domain factors AGL15 and AGL18 act redundantly as repressors of the floral transition in Arabidopsis. *Plant Journal* , *50* (6), 1007–1019. doi: 10.1111/j.1365-313X.2007.03105.x

Ansaldi, B. H., Weber, J. J., & Franks, S. J. (2018). The role of phenotypic plasticity and pollination environment in the cleistogamous, mixed mating breeding system of Triodanis perfoliata. *Plant Biology* , *20* (6), 1068–1074. doi: https://doi.org/10.1111/plb.12877

Arora, R., Agarwal, P., Ray, S., Singh, A. K., Singh, V. P., Tyagi, A. K., & Kapoor, S. (2007). MADS-box gene family in rice: Genome-wide identification, organization and expression profiling during reproductive development and stress. *BMC Genomics* , *8* . doi: 10.1186/1471-2164-8-242

Barchi, L., Rabanus-Wallace, M. T., Prohens, J., Toppino, L., Padmarasu, S., Portis, E., ... Giuliano, G. (2021). Improved genome assembly and pan-genome provide key insights into eggplant domestication and breeding. *The Plant Journal* , *107* (2), 579–596. doi: https://doi.org/10.1111/tpj.15313

Bartrina, I., Otto, E., Strnad, M., Werner, T., & Schmülling, T. (2011). Cytokinin regulates the activity of reproductive meristems, flower organ size, ovule formation, and thus seed yield in Arabidopsis thaliana. *Plant Cell* , *23* (1), 69–80. doi: 10.1105/tpc.110.079079

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research* , *27* (2), 573–580. doi: 10.1093/nar/27.2.573

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., ... Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* , *31* (1), 365–370. doi: 10.1093/nar/gkg095

Brahmachari, G., Mondal, S., Gangopadhyay, A., Gorai, D., Mukhopadhyay, B., Saha, S., & Brahmachari, A. K. (2004). Swertia (gentianaceae): Chemical and pharmacological aspects. *Chemistry and Biodiversity* ,*1* (11), 1627–1651. doi: 10.1002/cbdv.200490123

Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* ,*268* (1), 78–94. doi: 10.1006/jmbi.1997.0951

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., & Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology* ,*31* (12), 1119–1125. doi: 10.1038/nbt.2727

Campbell, C. S., Quinn, J. A., Cheplick, G. P., & Bell, T. J. (1983). Cleistogamy in grasses. *Annual Review of Ecology and Systematics. Vol. 14* , (39), 411–441. doi: 10.1146/annurev.es.14.110183.002211

Charlesworth, D., & Charlesworth, B. (1987). Inbreeding depression and its evolutionary consequences. *Annual Review of Ecology and Systematics. Vol. 18* , 237–268. doi: 10.1146/annurev.es.18.110187.001321

Chase, M. W., Christenhusz, M. J. M., Fay, M. F., Byng, J. W., Judd, W. S., Soltis, D. E., ... Weber, A. (2016). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society* , *181* (1), 1–20. doi: 10.1111/boj.12385

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* , *34* (17), i884–i890. doi: 10.1093/bioinformatics/bty560

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.*Bioinformatics* , *21* (18), 3674–3676.

Coordinators, N. R. (2016). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* ,*44* (D1), D7–D19. doi: 10.1093/nar/gkv1290

Culley, T. M., & Klooster, M. R. (2007). The cleistogamous breeding system: A review of its frequency, evolution, and ecology in angiosperms. *Botanical Review* , *73* (1), 1–30. doi: 10.1663/0006-8101(2007)73[1:TCBSAR]2.0.CO;2

Culley, T. M., & Wolfe, A. D. (2001). Population genetic structure of the cleistogamous plant species Viola pubescens Aiton (Violaceae), as indicated by allozyme and ISSR molecular markers. *Heredity* ,*86* (5), 545–556. doi: 10.1046/j.1365-2540.2001.00875.x

Darwin, C. (1897). *The different forms of flowers on plants of the same species* . D. Appleton.

De Bie, T., Cristianini, N., Demuth, J. P., & Hahn, M. W. (2006). CAFE: A computational tool for the study of gene family evolution.*Bioinformatics* , *22* (10), 1269–1271. doi: 10.1093/bioinformatics/btl097

Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., ... Lashermes, P. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis.*Science* , *345* (6201), 1181–1184. doi: 10.1126/science.1255274

Doyle, J. J., & Doyle, J. L. (1987). *A rapid DNA isolation procedure for small quantities of fresh leaf tissue* .

ES, C., & EM, M. (1991). The war of the whorls: genetic interactions controlling flower development. *Nature* , *353* (6339), 31.

Fawcett, M. M., Parks, M. C., Tibbetts, A. E., Swart, J. S., Richards, E. M., Vanegas, J. C., ... Angelini, D. R. (2018). Manipulation of insulin signaling phenocopies evolution of a host-associated polyphenism. *Nature Communications* , *9* (1), 1–11. doi: 10.1038/s41467-018-04102-1

Filiault, D. L., Ballerini, E. S., Mandáková, T., Aköz, G., Derieg, N. J., Schmutz, J., ... Nordborg, M. (2018). The Aquilegia genome provides insight into adaptive radiation and reveals an extraordinarily polymorphic chromosome with a unique history. *ELife* , *7* , e36426. doi: 10.7554/eLife.36426

Franke, J., Kim, J., Hamilton, J. P., Zhao, D., Pham, G. M., Wiegert-Rininger, K., ... O'Connor, S. E. (2019). Gene Discovery in Gelsemium Highlights Conserved Gene Clusters in Monoterpene Indole Alkaloid Biosynthesis. *ChemBioChem* , *20* (1), 83–87. doi: https://doi.org/10.1002/cbic.201800592

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... Zeng, Q. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data.*Nature Biotechnology* , *29* (7), 644.

Guo, W. J., Nagy, R., Chen, H. Y., Pfrunder, S., Yu, Y. C., Santelia, D., ... Martinoia, E. (2014). SWEET17, a facilitative transporter, mediates fructose transport across the tonoplast of arabidopsis roots and leaves. *Plant Physiology* , *164* (2), 777–789. doi: 10.1104/pp.113.232751

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., ... Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* , *9* (1), 1–22. doi: 10.1186/gb-2008-9-1-r7

He, T., Liu, S., & Liu, J. (2013). A new Qinghai-Tibet Plateau endemic genus Sinoswertia and its pollination mode. *Plant Diversity and Resources* , *35* (3), 393–400.

He, Z., Zhang, H., Gao, S., Lercher, M. J., Chen, W. H., & Hu, S. (2016). Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. *Nucleic Acids Research* , *44* (W1), W236–W241. doi: 10.1093/nar/gkw370

Hedges, S. B., Marin, J., Suleski, M., Paymer, M., & Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification.*Molecular Biology and Evolution* , *32* (4), 835–845. doi: 10.1093/molbev/msv037

Hoopes, G. M., Hamilton, J. P., Kim, J., Zhao, D., Wiegert-Rininger, K., Crisovan, E., & Buell, C. R. (2018). Genome Assembly and Annotation of the Medicinal Plant Calotropis gigantea, a Producer of Anticancer and Antimalarial Cardenolides. *G3 Genes|Genomes|Genetics* , *8* (2), 385–391. doi: 10.1534/g3.117.300331

Hu, J., Fan, J., Sun, Z., & Liu, S. (2020). NextPolish: A fast and efficient genome polishing tool for long-read assembly.*Bioinformatics* , *36* (7), 2253–2255. doi: 10.1093/bioinformatics/btz891

Hu, Y.-X., Tao, Y.-B., & Xu, Z.-F. (2017). Overexpression of Jatropha Gibberellin 2-oxidase 6 (JcGA2ox6) Induces Dwarfism and Smaller Leaves, Flowers and Fruits in Arabidopsis and Jatropha . *Frontiers in Plant Science* , Vol. 8, p. 2103. Retrieved from https://www.frontiersin.org/article/10.3389/fpls.2017.02103

Ishiguro, S., Kawai-Oda, A., Ueda, J., Nishida, I., & Okada, K. (2001). The defective in anther dehiscence1 gene encodes a novel phospholipase A1 catalyzing the initial step of jasmonic acid biosynthesis, which synchronizes pollen maturation, anther dehiscence, and flower opening in Arabidopsis. *Plant Cell* , *13* (10), 2191–2209. doi: 10.1105/tpc.13.10.2191

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., ... Characterization, T. F. P. C. for G. G. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* , *449* (7161), 463–467. doi: 10.1038/nature06148

Jibran, R., Tahir, J., Cooney, J., Hunter, D. A., & Dijkwel, P. P. (2017). Arabidopsis AGAMOUS regulates sepal senescence by driving jasmonate production. *Frontiers in Plant Science* ,*8* (December), 1–12. doi: 10.3389/fpls.2017.02101

Jin, J., Tian, F., Yang, D. C., Meng, Y. Q., Kong, L., Luo, J., & Gao, G. (2017). PlantTFDB 4.0: Toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Research* ,*45* (D1), D1040–D1045. doi: 10.1093/nar/gkw982

Joly, S., & Schoen, D. J. (2021). Repeated evolution of a reproductive polyphenism in plants is strongly associated with bilateral flower symmetry. *Current Biology* , 1–6. doi: 10.1016/j.cub.2021.01.009

Jurka, J., Kapitonov, V. V, Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* ,*110* (1–4), 462–467. doi: 10.1159/000084979

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods* , *14* (6), 587–589. doi:

10.1038/nmeth.4285

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability.*Molecular Biology and Evolution* , *30* (4), 772–780. doi: 10.1093/molbev/mst010

Kaul, S., Koo, H. L., Jenkins, J., Rizzo, M., Rooney, T., Tallon, L. J., ... Lin, X. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* , *408* (6814), 796–815.

Ke, M., Gao, Z., Chen, J., Qiu, Y., Zhang, L., & Chen, X. (2018). Auxin controls circadian flower opening and closure in the waterlily.*BMC Plant Biology* , *18* (1), 143. doi: 10.1186/s12870-018-1357-7

Keilwagen, J., Hartung, F., & Grau, J. (2019). GeMoMa: Homology-based gene prediction utilizing intron position conservation and RNA-seq data. In *Methods in Molecular Biology* (Vol. 1962, pp. 161–177). doi: 10.1007/978-1-4939-9173-0_9

Kellner, F., Kim, J., Clavijo, B. J., Hamilton, J. P., Childs, K. L., Vaillancourt, B., ... O'Connor, S. E. (2015). Genome-guided investigation of plant natural product biosynthesis. *The Plant Journal* , *82* (4), 680–692. doi: https://doi.org/10.1111/tpj.12827

Kellner, F., Kim, J., Clavijo, B. J., Hamilton, J. P., Childs, K. L., Vaillancourt, B., ... O'Connor, S. E. (2015). Genome-guided investigation of plant natural product biosynthesis. *Plant Journal* , *82* (4), 680–692. doi: 10.1111/tpj.12827

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods* ,*12* (4), 357–360. doi: 10.1038/nmeth.3317

Kim, S.-Y., Yu, H.-J., Hong, J. K., Woo, J. G., & Ahn, Y. K. (2013). Functional analysis of female gametophyte specific promoters in Chinese cabbage. *Scientia Horticulturae* , *156* , 29–37. doi: https://doi.org/10.1016/j.scienta.2013.03.003

Klemens, P. A. W., Patzke, K., Deitmer, J., Spinner, L., Le Hir, R., Bellini, C., ... Ekkehard Neuhaus, H. (2013). Overexpression of the vacuolar sugar carrier AtSWEET16 modifies germination, growth, and stress tolerance in Arabidopsis. *Plant Physiology* , *163* (3), 1338–1352. doi: 10.1104/pp.113.224972

Koontz, S. M., Weekley, C. W., Haller Crate, S. J., & Menges, E. S. (2017). Patterns of chasmogamy and cleistogamy, a mixed-mating strategy in an endangered perennial. *AoB Plants* , *9* (6), plx059.

Kram, B. W., & Carter, C. J. (2009). Arabidopsis thaliana as a model for functional nectary analysis. *Sexual Plant Reproduction* ,*22* (4), 235–246. doi: 10.1007/s00497-009-0112-5

Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* , *9* . doi: 10.1186/1471-2105-9-559

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* , *9* (4), 357–359. doi: 10.1038/nmeth.1923

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM* . *00* (00), 1–3. Retrieved from http://arxiv.org/abs/1303.3997

Li, H. T., Yi, T. S., Gao, L. M., Ma, P. F., Zhang, T., Yang, J. B., ... Li, D. Z. (2019). Origin of angiosperms and the puzzle of the Jurassic gap. *Nature Plants* , *5* (5), 461–470. doi: 10.1038/s41477-019-0421-0

Li, L., Stoeckert, C. J. J., & Roos, D. S. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes – Li et al. 13 (9): 2178 – Genome Research. *Genome Research* , *13* (9), 2178–2189. doi: 10.1101/gr.1224503.candidates

Lin, I. W., Sosso, D., Chen, L.-Q., Gase, K., Kim, S.-G., Kessler, D., ... Frommer, W. B. (2014). Nectar secretion requires sucrose phosphate synthases and the sugar transporter SWEET9. *Nature* ,*508* (7497), 546–549. doi: 10.1038/nature13082

Liu, Yiyang, Zhang, X., Han, K., Li, R., Xu, G., Han, Y., ... Wan, S. (2021). Insights into amphicarpy from the compact genome of the legume Amphicarpaea edgeworthii. *Plant Biotechnology Journal* ,*19* (5), 952–965. doi: 10.1111/pbi.13520

Liu, Yue, Wang, Y., Guo, F., Zhan, L., Mohr, T., Cheng, P., ... Gu, Y. Q. (2017). Deep sequencing and transcriptome analyses to identify genes involved in secoiridoid biosynthesis in the Tibetan medicinal plant Swertia mussotii. *Scientific Reports* , *7* (September 2016), 1–14. doi: 10.1038/srep43108

Lord, E. M. (1981). Cleistogamy: A tool for the study of floral morphogenesis, function and evolution. *The Botanical Review* ,*47* (4), 421–449. doi: 10.1007/BF02860538

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* , *15* (12), 1–21. doi: 10.1186/s13059-014-0550-8

Lu, T., Ke, M., Lavoie, M., Jin, Y., Fan, X., Zhang, Z., ... Zhu, Y.-G. (2018). Rhizosphere microorganisms can influence the timing of plant flowering. *Microbiome* , *6* (1), 231. doi: 10.1186/s40168-018-0615-0

Majoros, W. H., Pertea, M., & Salzberg, S. L. (2004). TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders.*Bioinformatics* , *20* (16), 2878–2879. doi: 10.1093/bioinformatics/bth315

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.*Bioinformatics* , *27* (6), 764–770. doi: 10.1093/bioinformatics/btr011

Mayr, E. (1963). *Animal species and evolution* . Belknap Press of Harvard University Press,.

Moran, N. A. (1992). The evolutionary maintenance of alternative phenotypes. *American Naturalist* , *139* (5), 971–989. doi: 10.1086/285369

Morinaga, S.-I., Nagano, A. J., Miyazaki, S., Kubo, M., Demura, T., Fukuda, H., ... Hasebe, M. (2008). Ecogenomics of cleistogamous and chasmogamous flowering: genome-wide gene expression patterns from cross-species microarray analysis in Cardamine kokaiensis (Brassicaceae). *Journal of Ecology* , *96* (5), 1086–1097. doi: https://doi.org/10.1111/j.1365-2745.2008.01392.x

Nagpal, P., Ellis, C. M., Weber, H., Ploense, S. E., Barkawi, L. S., Guilfoyle, T. J., ... Reed, J. W. (2005). Auxin response factors ARF6 and ARF8 promote jasmonic acid production and flower maturation.*Development* , *132* (18), 4107–4118. doi: 10.1242/dev.01955

Ng, M., & Yanofsky, M. F. (2001). Function and evolution of the plant MADS-box gene family. *Nature Reviews Genetics* , *2* (3), 186–195. doi: 10.1038/35056041

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* ,*32* (1), 268–274. doi: 10.1093/molbev/msu300

Organization, W. H. (2002). *Traditional medicine in asia* . WHO Regional Office for South-East Asia.

Oudin, A., Courtois, M., Rideau, M., & Clastre, M. (2007). The iridoid pathway in Catharanthus roseus alkaloid biosynthesis.*Phytochemistry Reviews* , *6* (2–3), 259–276. doi: 10.1007/s11101-006-9054-9

Parenicova, L., de Folter, S., Kieffer, M., Horner, D. S., Favalli, C., Busscher, J., ... Colombo, L. (2003). Molecular and Phylogenetic Analyses of the Complete MADS-Box Transcription Factor Family in Arabidopsis : New Openings to the MADS World[W]. *The Plant Cell* , *15* (7), 1538–1551. doi: 10.1105/tpc.011544

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* ,*33* (3), 290–295. doi: 10.1038/nbt.3122

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., & Lopez, R. (2005). InterProScan: Protein domains identifier. *Nucleic Acids Research* , *33* (SUPPL. 2), 116–120. doi: 10.1093/nar/gki442

Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* , *11* (1). doi: 10.1038/s41467-020-14998-3

Rao, M. U., Sreenivasulu, M., Chengaiah, B., Reddy, K. J., & Chetty, C. M. (2010). Herbal medicines for diabetes mellitus: a review. *Int J PharmTech Res* , *2* (3), 1883–1892.

Rédei, G. P. (Ed.). (2008). *BLASTP BT - Encyclopedia of Genetics, Genomics, Proteomics and Informatics* . Dordrecht: Springer Netherlands. doi: 10.1007/978-1-4020-6754-9_1881

Schnee, B. K., & Waller, D. M. (1986). Reproductive Behavior of Amphicarpaea Bracteata (Leguminosae), an Amphicarpic Annual.*American Journal of Botany* , *73* (3), 376–386. doi: 10.1002/j.1537-2197.1986.tb12051.x

Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C. J., Vert, J. P., ... Barillot, E. (2015). HiC-Pro: An optimized and flexible pipeline for Hi-C data processing. *Genome Biology* , *16* (1), 1–11. doi: 10.1186/s13059-015-0831-x

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* ,*31* (19), 3210–3212. doi: 10.1093/bioinformatics/btv351

Simpson, S. J., Sword, G. A., & Lo, N. (2011). Polyphenism in insects.*Current Biology* , *21* (18), R738–R749. doi: 10.1016/j.cub.2011.06.006

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., & Ideker, T. (2011). Cytoscape 2.8: New features for data integration and network visualization. *Bioinformatics* , *27* (3), 431–432. doi: 10.1093/bioinformatics/btq675

Stanke, M., Steinkamp, R., Waack, S., & Morgenstern, B. (2004). AUGUSTUS: A web server for gene finding in eukaryotes. *Nucleic Acids Research* , *32* (WEB SERVER ISS.). doi: 10.1093/nar/gkh379

Sun, P., Jiao, B., Yang, Y., Shan, L., Li, T., Li, X., ... Liu, J. (2021). WGDI: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *BioRxiv* .

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments.*Nucleic Acids Research* , *34* (WEB. SERV. ISS.), 609–612. doi: 10.1093/nar/gkl315

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* . doi: 10.1002/0471250953.bi0410s25

Teotia, S., & Tang, G. (2015). To Bloom or Not to Bloom: Role of MicroRNAs in Plant Flowering. *Molecular Plant* , *8* (3), 359–377. doi: https://doi.org/10.1016/j.molp.2014.12.018

Thiel, T., Michalek, W., Varshney, R. K., & Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.).*Theoretical and Applied Genetics* , *106* (3), 411–422. doi: 10.1007/s00122-002-1031-0

Tian, F., Yang, D. C., Meng, Y. Q., Jin, J., & Gao, G. (2020). PlantRegMap: Charting functional regulatory maps in plants.*Nucleic Acids Research* , *48* (D1), D1104–D1113. doi: 10.1093/nar/gkz1020

Torti, S., Fornara, F., Vincent, C., Andrés, F., Nordström, K., Göbel, U., ... Coupland, G. (2012). Analysis of the Arabidopsis shoot meristem transcriptome during floral transition identifies distinct regulatory patterns and a leucine-rich repeat protein that promotes flowering. *Plant Cell* , *24* (2), 444–462. doi: 10.1105/tpc.111.092791

van Tunen, A. J., Eikelboom, W., & Angenent, G. C. (1993). Floral organogenesis in Tulipa. *Flowering Newsletter* , (16), 33–38.

Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads.*Genome Research* , *27* (5), 737–746. doi: 10.1101/gr.214270.116

Vranová, E., Coman, D., & Gruissem, W. (2013). Network analysis of the MVA and MEP pathways for isoprenoid synthesis. *Annual Review of Plant Biology* , *64* , 665–700. doi: 10.1146/annurev-arplant-050312-120116

Waller, D. M. (1980). Environmental Determinants of Outcrossing in Impatiens capensis (Balsaminaceae). *Evolution* , *34* (4), 747–761. doi: 10.2307/2408029

Wang, C. H., Du, W., & Wang, X. F. (2017). Reproductive investment in a cleistogamous morph of Polygonum jucundum (Polygonaceae). *Plant Systematics and Evolution* , *303* (4), 559–563. doi: 10.1007/s00606-017-1388-9

Wuyun, T., Wang, L., Liu, H., Wang, X., Zhang, L., Bennetzen, J. L., ... Du, H. (2018). The Hardy Rubber Tree Genome Provides Insights into the Evolution of Polyisoprene Biosynthesis. *Molecular Plant* ,*11* (3), 429–442. doi: https://doi.org/10.1016/j.molp.2017.11.014

Xia, E., Tong, W., Hou, Y., An, Y., Chen, L., Wu, Q., ... Wan, X. (2020). The Reference Genome of Tea Plant and Resequencing of 81 Diverse Accessions Provide Insights into Its Genome Evolution and Adaptation.*Molecular Plant* , *13* (7), 1013–1026. doi: https://doi.org/10.1016/j.molp.2020.04.010

Xu, Z., Xu, Z., Pu, X., Gao, R., Demurtas, O. C., Fleck, S. J., ... Song, J. (2020). Tandem gene duplications drive divergent evolution of caffeine and crocin biosynthetic pathways in plants. *BMC Biology* ,*18* (1), 1–14. doi: 10.1186/s12915-020-00795-3

Yang, C.-H., & Andrew Pospisilik, J. (2019). Polyphenism – A Window Into Gene-Environment Interactions and Phenotypic Plasticity .*Frontiers in Genetics* , Vol. 10, p. 132. Retrieved from https://www.frontiersin.org/article/10.3389/fgene.2019.00132

Yang, L., Zhou, G., & Chen, G. (2011). Genetic diversity and population structure of Swertia tetraptera (Gentianaceae), an endemic species of Qinghai-Tibetan Plateau. *Biochemical Systematics and Ecology* ,*39* (4–6), 302–308. doi: 10.1016/j.bse.2011.08.003

Yang, Y., Sun, P., Lv, L., Wang, D., Ru, D., Li, Y., ... Liu, J. (2020). Prickly waterlily and rigid hornwort genomes shed light on early angiosperm evolution. *Nature Plants* , *6* (3), 215–222. doi: 10.1038/s41477-020-0594-6

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood.*Molecular Biology and Evolution* , *24* (8), 1586–1591. doi: 10.1093/molbev/msm088

Yu, G., Wang, L. G., Han, Y., & He, Q. Y. (2012). ClusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS A Journal of Integrative Biology* , *16* (5), 284–287. doi: 10.1089/omi.2011.0118

Yu, J., Hu, S., Wang, J., Wong, G. K. S., Li, S., Liu, B., ... Yang, H. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science* , *296* (5565), 79–92. doi: 10.1126/science.1068037

Yu, X., Xiao, J., Chen, S., Yu, Y., Ma, J., Lin, Y., ... Liu, R. (2020). Metabolite signatures of diverse Camellia sinensis tea populations. *Nature Communications* , *11* (1), 5586. doi: 10.1038/s41467-020-19441-1

Zhang, J., Wu, F., Yan, Q., John, U. P., Cao, M., Xu, P., ... Wang, Y. (2021). The genome of Cleistogenes songorica provides a blueprint for functional dissection of dimorphic flower differentiation and drought adaptability. In *Plant Biotechnology Journal* (Vol. 19). doi: 10.1111/pbi.13483

Zheng, Q., Zheng, Y., Ji, H., Burnie, W., & Perry, S. E. (2016). Gene regulation by the AGL15 transcription factor reveals hormone interactions in somatic embryogenesis. *Plant Physiology* ,*172* (4), 2374–2387. doi: 10.1104/pp.16.00564

Figure Legends

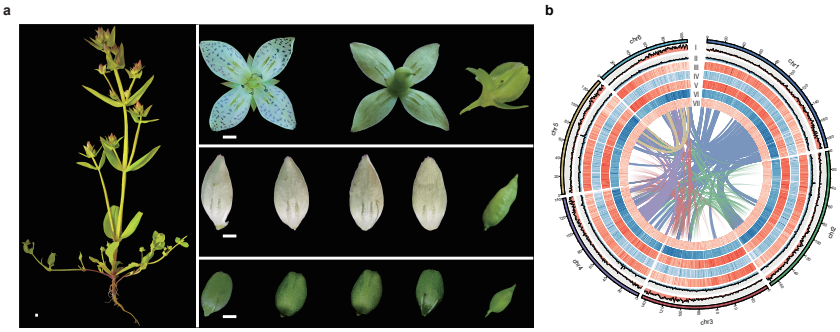
Figure 1. Overview of the morphology and genome assembly of*Sinoswertia tetraptera.* (a) Morphology of the chasmogamous (CH) and cleistogamous (CL) flowers and petals. Scale bars, 1 cm. (b) Genome assembly and annotations. From inside to outside: (I) gene density in 500kb sliding windows, (II) GC density in 500kb sliding windows, (III) Gypsy density in 500kb sliding windows, (IV) Copia density in 500kb sliding windows, (V) density of long terminal repeats (LTR) in 500kb sliding windows, (VI) density of TE in 500kb sliding windows. (VIII) the density of SNP in 50kb sliding windows.

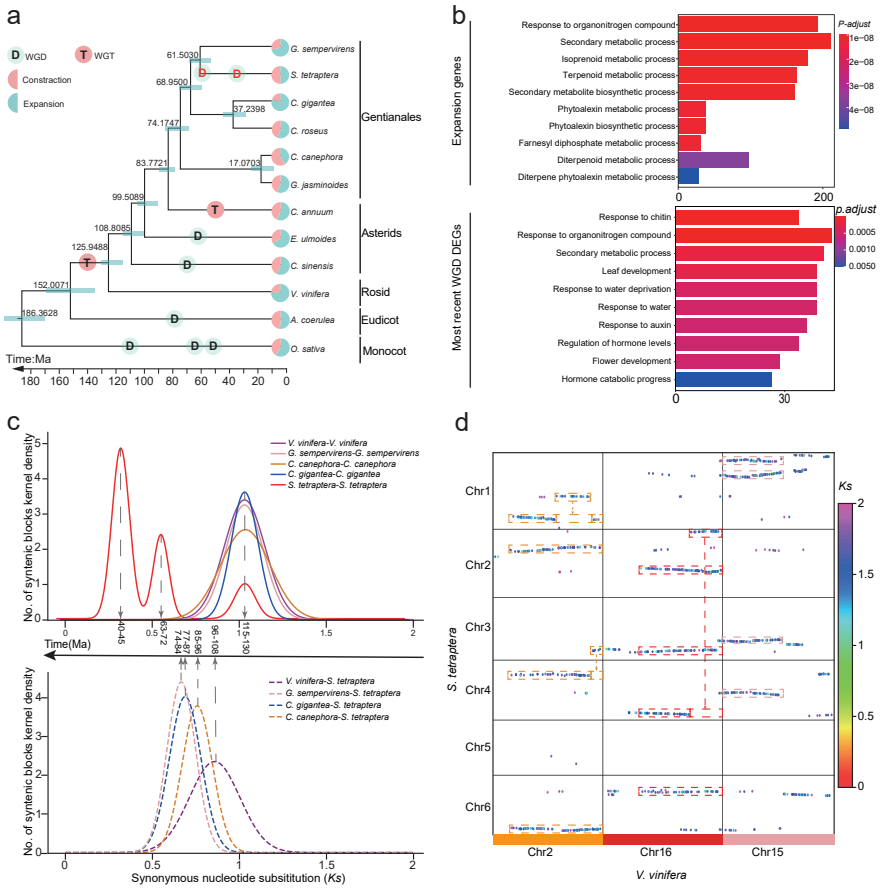
Figure 2. Genome evolution. (a) Chronogram showing divergence times and genome duplications in angiosperms (Gentianales, Asterids, Rosid, Eudicot, and Monocot), with node age and 95% confidence intervals. The lengths of light blue labels represent the random values in Ma. Dots with * represent resolved polyploidization events in previous studies; others indicate new events we identified in this study, while D indicates duplication events and T triplication events. Pie charts show the proportions of gene families among the 12 species that underwent expansion or contraction. (b) Functional enrichment analysis of genes belonging to the last WGD which different expressed between dimorphic flowers (hear called most recent WGD DEGs) and expanded in *S. tetraptera* . The length of bars represents the number of genes. The enriched GO terms of biological progress with corrected P -value <0.01 are presented. Terms presented after clusterProfiler simplify. (c) After evolutionary rate correction among the various species, the distribution of average synonymous substitution levels (*Ks*) between syntenic blocks was raised by different color lines for each species. (I) *Ks* distribution showing *Ks* distribution from paralogs within a species. (II)*Ks* distribution showing *Ks* from orthologs between *S. tetraptera* and each of four species indicated by dashed lines. (d) Synthetic blocks (involving [?] 10 colinear genes) between genomes involving *S. tetraptera* and *V. vinifera.* The corresponding median *Ks* value is shown for each block, and the various colored rectangles represent polyploidization events. The homologous chromosomes in grape were selected and are presented in blue.

Figure 3. MADS-box genes and genetic regulation of closed and open dimorphic flowers of *S. tetraptera* **.** (a) A phylogenetic tree of the MADS-box gene family. The numbers of the *AGL15/18* sub-class members within five species and their significance were obtained by the LSD test after Bonferroni (BH) correction. (b) The level of plant hormones participating in bloom regulation in CH and CL flowers. **p < 0.01, Student's t-test. (c) A proposed pathway for the control of closed or open dimorphic flowers. Gene expression profiles are presented in the heatmap alongside the gene names. The bar shows the expression level of each gene. Low to high expression is indicated by light yellow to red.

Figure 4. Differentiation between CH and CL flowers. (a) The floral size and the level of cytokinin differ between CH and CL flowers. **p < 0.01, Student's t-test. The lower part of the figure presents a probable cytokinin-dependent pathway controlling the contrasting sizes of dimorphic flowers. Gene expression profiles are presented in the heatmap alongside the gene names. (b) Gene expression profiles involved in flower pigment biosynthesis of CH and CL flowers. (c) Gene expression profiles of the *SWEET* gene family members, comparing CH and CL flowers.

Figure 5. Iridoid biosynthesis in *S. tetraptera*. (a) Concentration of four iridoids in seven tissues. Different letters within each part indicate significance according to the LSD test after Bonferroni (BH) correction. (b) Sub-network for the leaf and flower module of iridoid biosynthesis. (c) The probable biosynthesis pathway of four iridoids. Gene expression profiles are presented in the heatmap alongside the gene names.





Hosted file

Figure_3.pdf available at https://authorea.com/users/448888/articles/547583-multi-omics-reveal-differentiation-and-maintenance-of-dimorphic-flowers-in-an-alpine-plant-on-the-qinghai-tibet-plateau