New Phytologist

Research

The genome of a cave plant, *Primulina huaijiensis*, provides insights into adaptation to limestone karst habitats

Chao Feng[1], Jing Wang[1], Lingqing Wu[2], Hanghui Kong[1], Lihua Yang[1], Chen Feng[1], Kai Wang[2], Mark Rausher[3] and Ming Kang[1,4]

[1]Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, China; [2]Novogene Bioinformatics Institute, Beijing 100083, China; [3]Department of Biology, Duke University, 125 Science Drive, Durham, NC 27705, USA; [4]Center of Conservation Biology, Core Botanical Gardens, Chinese Academy of Sciences, Guangzhou 510650, China

Authors for correspondence:
Ming Kang
Tel: +86 20 37082193
Email: mingkang@scbg.ac.cn

Mark Rausher
Tel: +1 919 684 2295
Email: mrausher@duke.edu

Kai Wang
Tel: +86 13299910027
Email: wangkai@novogene.cn

Received: *25 November 2019*
Accepted: *29 March 2020*

New Phytologist (2020) **227:** 1249–1263
doi: 10.1111/nph.16588

Key words: chromosome-level genome, differential gene retention, gene family expansion, karst habitat adaptation, *Primulina huaijiensis*, whole genome duplication.

Summary

- Although whole genome duplication (WGD) has been suggested to facilitate adaptive evolution and diversification, the role of specific WGD events in promoting diversification and adaptation in angiosperms remains poorly understood. *Primulina*, a species-rich genus with > 180 species associated with limestone karst habitat, constitutes an ideal system for studying the impact of WGD events on speciation and evolutionary adaptation.
- We sequenced and assembled a chromosome-level genome of the cave-dwelling species *P. huaijiensis* to study gene family expansion and gene retention following WGDs.
- We provide evidence that *P. huaijiensis* has undergone two WGDs since the γ triplication event shared by all eudicots. In addition to a WGD shared by almost all Lamiales (*L* event), we identified a lineage-specific WGD (*D* event) that occurred in the early Miocene around 20.6–24.2 Myr ago and that is shared by almost the entire subtribe Didymocarpinae. We found that gene retentions following the *D* event led to gene family proliferation (e.g. WRKYs) that probably facilitated adaptation to the high salinity and drought stress in limestone karst.
- Our study highlights the role of lineage-specific WGD in species diversification and adaptation of plants from special habitats.

Introduction

Limestone karst caves have played a key role in understanding adaptation and speciation ever since Wallace's work in the Malay Archipelago (Wallace, 1858). Cave-dwelling plants represent an attractive system for studying evolution in extreme environments because, like most plants, their sessile nature forces them to directly cope with environmental conditions rather than escape to more favorable sites. Cave and cave-like habitats associated with limestone karsts are characterized by abiotic and biotic properties that may constitute strong selective agents for a wide range of plant traits. These include low light intensity, pollinator limitation, shallow soil with low water-holding capacity, and reduced availability of essential plant nutrients (Hao *et al.*, 2015). These intense selective pressures, coupled with the patchy, island-like distribution of cave-like habitats, is likely to foster the evolution of high species endemism (Monro *et al.*, 2018). Such endemism is apparent in Southeast Asia and southern China, where limestone karst landforms have generated abundant cave and cave-like habitats. These habitats support a remarkably high level of species endemism in flowering plants (Wei, 2010; Chung *et al.*, 2014; Monro *et al.*, 2018). Although species diversity of individual caves is usually low, plant composition varies greatly from one cave to another, leading to karst landforms being recognized as 'natural laboratories' for evolutionary studies (Clements *et al.*, 2006; Oliver *et al.*, 2017). However, the precise nature of the adaptive changes and speciation processes that have generated this biodiversity remain poorly understood (Wang *et al.*, 2017a,b).

Whole genome duplication (WGD), or polyploidy, has long been recognized as a prominent process facilitating adaptive evolution and diversification (Doyle *et al.*, 2008; Soltis *et al.*, 2009; Wu *et al.*, 2020). Recent genomic analyses suggest that all angiosperms have undergone at least two rounds of shared WGD during their evolutionary history (Cui *et al.*, 2006; Soltis *et al.*, 2009; Jiao

et al., 2011). However, the role of WGD events in promoting diversification in angiosperms remains poorly understood, with recent studies providing conflicting results about the relationship between WGD and shifts in diversification rates (Madlung, 2013; Landis *et al.*, 2018). For example, Edger *et al.* (2015) found increased diversification rates following ancient WGDs in Brassicales. Using a 639-taxon time-calibrated tree representing angiosperm phylogeny and nine WGD events, Tank *et al.* (2015) found that at least half of the WGD events investigated had an impact on diversification. This hypothesis was further supported by a recent analysis using a larger phylogeny and many more WGD events (Landis *et al.*, 2018). By contrast, there was no direct association between WGD events and increased diversification in a recent analysis of Caryophyllales (Smith *et al.*, 2018).

A WGD duplicates all of the nuclear genes of an organism simultaneously, which provides novel genetic material that may facilitate adaptation and promote speciation (Hegarty & Hiscock, 2008; Van de Peer *et al.*, 2009; Jiao *et al.*, 2011). Recent studies have indicated that after WGD events, species tend to retain a large fraction of duplicates with specific molecular functions, leading to increases in the sizes of gene families and creating the opportunity for duplicate copies to participate in lineage-specific adaptive change (Ren *et al.*, 2018; Wu *et al.*, 2020). Many flowering plants have undergone multiple rounds of WGDs (Vision *et al.*, 2010; Geiser *et al.*, 2016; Wang *et al.*, 2019), resulting in hundreds to thousands of retained gene duplicates, with possible differential retention or loss of gene duplicates among different lineages. However, the contributions of different WGDs to gene family expansions and gene retention remain poorly explored.

Primulina (Gesneriaceae) is a monophyletic genus with > 180 described species that are distributed widely across the limestone karsts of southern China and northern Vietnam (Xu *et al.*, 2017). They represent a group of typical 'stone plants' that have adapted to remarkably diverse habitats and niches, from cave and cave-like habitats to steep cliffs. However, almost all species exhibit edaphic specialization, with the majority occurring in calcareous soils developed from karst limestone bedrock (i.e. calciphiles), but with a few growing solely on acid soils (i.e. calciphobes) (Hao *et al.*, 2015). Presumably due to the terrestrial-island nature of karst landforms in southern China (Gao *et al.*, 2015), most species are micro-endemics with narrow distributions, often limited to a single cave or limestone hill system. The high species richness and endemism of the genus, together with the high degree of habitat specialization make *Primulina* an excellent model for studying evolutionary adaptation to karst habitat environments.

Recent investigations show that the availability of genomic tools greatly facilitates the elucidation of the processes responsible for adaptive divergence between species (Ellegren, 2014). Because there is no published genome for any *Primulina* species, we have been limited in our abilities to fully understand the processes responsible for diversification and endemism in this genus. We therefore have undertaken the sequencing and assembly of the genome of the cave-dwelling species *P. huaijiensis* ($2n = 2x = 36$) (Kang *et al.*, 2014). Here we present a chromosome-level genome and use it as a basis to infer characteristics of the evolutionary radiation of the genus. We demonstrate that a lineage-specific WGD and consequent gene family expansions may have facilitated species diversification and adaptation in *Primulina* in karst cave habitats.

Materials and Methods

Genome sequencing, assembly and characterization

Primulina huaijiensis is a micro-endemic restricted to a limestone karst cave in northwest Guangdong and has the smallest genome size (*c.* 547 Mbp) in the genus *Primulina* (Kang *et al.*, 2014). Several individuals were introduced and cultivated at the South China Botanical Garden (SCBG), Chinese Academy of Sciences (CAS) (Guangzhou, China). We extracted the genomic DNA from fresh leaves using a modified CTAB method (Doyle, 1990). We constructed six paired-end libraries with short insert sizes of 230, 350 and 500 bp, and eight mate-pair libraries with insert sizes of 2, 5, 10 and 15 kbp. These libraries were subjected to paired-end (PE) 125/150 bp sequencing on HiSeq 2500/HiSeq X Ten platform (Supporting Information Table S1). We filtered the raw data by removing PCR duplications, adapter sequences and low-quality sequences with < 90% identified nucleotides using our previous in-house pipeline QC_pe (https://github.com/scbgfengchao/; Figshare doi: 10.6084/m9.figshare.10185056; Feng *et al.*, 2017). For mate-pair libraries, we also used DeLoxer (http://genomes.ucsd.edu/downloads) to remove the unpaired reads. After estimating the genome size, heterozygosity, repeat rate of *P. huaijiensis* by the k-mer method using Geo software (ftp://ftp.genomics.org.cn/pub/gce), we assembled the *P. huaijiensis* genome according to a hybrid-specific SOAPdenovo approach (Huang *et al.*, 2016; S. Wang *et al.*, 2017; Wan *et al.*, 2018). Further, we prepared a Hi-C library following the standard procedure (Lieberman-Aiden *et al.*, 2009). After mapping against the primary scaffolds using Bwa (Li & Durbin, 2009), we corrected, clustered, sorted and anchored the scaffolds with the length over 1 kbp into 18 pseudo-molecules using Achesis (Burton *et al.*, 2013). To evaluate the consistency and completeness of the assembly, we carried out a comprehensive analysis that included constructing a heat map of chromosome interactions, a 2D surface distribution of GC content and sequencing depth, short-insert library read mapping, Core Eukaryotic Genes (CEG) alignment, Benchmarking Universal Single-copy Orthologs (BUSCO) alignment, EST alignment and RNA-Seq read mapping.

We identified repetitive sequences at the DNA and protein levels by a combination of homology-based prediction and *de novo* identification. We predicted protein-coding gene structures by a combination of *de novo* identification, homology-based prediction and RNA-Seq-based prediction, and then integrated this information into a nonredundant gene model set by using EVM (Haas *et al.*, 2008). Additionally, we annotated the protein-coding genes against Swissprot, TrEMBL, KEGG and InterPro databases. We identified tandem duplications (i.e. tandemly repeated gene arrays) using our in-house script TD_identification

(https://github.com/scbgfengchao/ and Figshare: 10.6084/m9.figshare.10185056), tolerating one unrelated gene among cluster members. See Methods S1 for additional details on assembly, evaluation and annotation.

Orthogroup clustering and comparative phylogeny analysis across angiosperm species

We used OrthoFinder v.2.3.3 (Emms & Kelly, 2015) with the parameter (-S diamond -og) to classify the orthogroups of proteins from *P. huaijiensis* and 16 other model sequenced plants, including *Tectona grandis*, *Handroanthus impetiginosus*, *Sesamum indicum*, *Antirrhimum majus*, *Dorcoceras hygrometricum* (original name: *Boea hygrometrica*; Puglisi *et al.*, 2016), *Olea europaea*, *Fraxinus excelsior*, *Solanum tuberosum*, *Solanum lycopersicum*, *Arabidopsis thaliana*, *Theobroma cacao*, *Populus trichocarpa*, *Vitis vinifera*, *Oryza sativa*, *Musa acuminata* and *Amborella trichopoda*.

For phylogeny construction, we selected proteins of single-copy orthogroups (i.e. the orthogroups that contain only one or none genes for each species) presented in ≥ 70% of species, and aligned them using Mafft (v.6.864b) (Katoh & Standley, 2013). We then converted them into aligned coding sequences (CDS) using Pal2nal script (Suyama *et al.*, 2006). After determination of the best substitution model for each orthogroup using IQ-tree (v.1.7-beta12) (Nguyen *et al.*, 2015) and discarding the orthogroups with partition-specific rates > 2.0 or < 0.5, we constructed the maximum-likelihood (ML) phylogenetic trees across the 17 plant species using IQ-tree with the parameter (-p -bb 1000), setting *A. trichopoda* as outgroup.

For species divergence time estimation, we applied the software r8s v.1.83 (Sanderson, 2003) with the parameter 'smooth' of 1, setting two fossil constraints (stem group of Brassicales, stem group of *Fraxinus*) and a secondary calibration node (the ancestor node of eudicots and monocots; 177.10 Myr ago (Ma); Foster *et al.*, 2017). For the basis for assigning these fossils to the calibrated nodes, we followed Li *et al.* (2019) in placing the fossil of *Dressiantha bicarpelata* (age: Turonian, 89.8 Ma; Gandolfo *et al.*, 1998) at the stem group of Brassicales (i.e. the ancestor node of *A. thaliana* and *T. cacao*), and followed Roalson & Roberts (2016) in placing the fossil of *Fraxinus wilcoxiana* (age: Middle Eocene, 44.3 Ma; Call & Dilcher, 1992) at the stem group of *Fraxinus* (i.e. the ancestor node of *F. excelsior* and *O. europaea*), respectively. We calculated the 95% confidence interval for fossil dates using our in-house scripts (r8s_CI, https://github.com/scbgfengchao/; Figshare, doi: 10.6084/m9.figshare.10185056). First, we generated 2000 bootstrap samples of proteins by randomly selecting 5% of the single-copy orthogroups. Then we constructed the ML phylogenetic trees, and filtered out the ones inconsistent with the topology of the known tree based on all the single-copy orthogroups. Finally, we calculated the divergence time of each remaining tree using r8s. For gene family expansion analysis, we investigated the ancestral gene content of each cluster at each node using Cafe v.3.1 (De Bie *et al.*, 2006) on a basis of phylogeny and gene numbers per orthogroup in each species, and then determined the gene family expansions or contractions at each branch with *P*-value < 0.01.

Transcriptome assembly

We obtained transcriptomes from eight subtribe Didymocarpinae species, including *Henckelia anachoreta*, *Cyrtandra dispar*, *Hemiboea subcapitata*, *Petrocodon fangianus*, *Primulina rubella*, *Primulina swinglei*, *Primulina fimbrisepala* and *Primulina eburnea*. The first five were newly sequenced in this study, whereas the latter three were obtained from our previous study (Ai *et al.*, 2015). The four *Primulina* species span the four main clades of the genus (Kong *et al.*, 2017). After filtering the raw data, we assembled the reads using Trinity v.2.4.0 (Grabherr *et al.*, 2011). Then we used the longest isoform from each Trinity assembly to generate unigene by using our in-house script (Trinity2Unigene.pl, https://github.com/scbgfengchao/; Figshare doi: 10.6084/m9.figshare.10185056), and further reduced the redundancy of unigenes using Cd-hit-est v.4.7 (with the parameter -c 0.98) (Fu *et al.*, 2012). After that, we identified the coding regions (cds and protein sequences) of each species by using TransDecoder (Haas *et al.*, 2013). To evaluate the completeness of the genes, we carried out Busco alignment against lineage dataset embryophyta_odb10.

Divergence time estimation and ortholog inference across asterids

We clustered the orthogroups from two datasets by using OrthoFinder v.2.3.3 (Emms & Kelly, 2015). In Dataset 1, we selected proteins from nine genomes (*T. grandis*, *H. impetiginosus*, *S. indicum*, *A. majus*, *P. huaijiensis*, *D. hygrometricum*, *O. europaea*, *F. excelsior* and *S. lycopersicum*) and eight Didymocarpinae transcriptomes. In Dataset 2, we added two species (*S. lycopersicum* and *V. vinifera*) to Dataset 1.

We selected Dataset 2 to construct the phylogeny following the pipeline as stated above, and then to estimate the divergence time of the 19 species across asterids by r8s v.1.83 (Sanderson, 2003), setting one fossil constraint (the stem group for *Fraxinus*, 44.3 Ma) and three secondary calibrations (the divergence time between grape and asterids, the crown age of Lamiales, and the crown age of *Primulina*). The first two secondary calibrations were obtained from the estimated divergence time of the 17 species across angiosperm in this study, whereas the last one was referenced from our previous work (14.14 Ma) on the phylogeny of *Primulina* genus covering approx. 160 species (Kong *et al.*, 2017). Also, on the basis of the r8s result, we obtained the substitution rate of each node of 19 species and their ancestors.

We identified homologs and orthologs from both Datasets 1 and 2 following the pipeline of Yang & Smith (2014) (https://bitbucket.org/yangya/phylogenomic_dataset_construction). Initially, we aligned each orthogroup presented in all of the species (i.e. the orthogroups that contain one or more genes for each species) using Mafft (v.6.864b) (Katoh & Standley, 2013), inferred two round ML phylogenies using RAxML (v.8.2.9) (Stamatakis, 2014), and trimmed those phylogenies with tips both longer than 0.2 and over 10 times longer than the average distance to tips of its sister clade, and also discarded those with branches longer than 0.5. According to the definition of Yang & Smith (2014) and Yang *et al.* (2015), the remaining trees were homolog trees.

Further, we obtained the 1 : 1 orthologs presented in all of the ingroup species (i.e. those orthologs that contain only one gene for each species) by pruning the homolog trees using the RT (rooted ingroups) method (i.e. prune by extracting ingroup clades and then cut paralogs from root to tip) (Yang & Smith, 2014), with full taxon occupancy (i.e. the pruned trees should contain outgroup and all of ingroup taxa).

Lastly, we calculated the synonymous substitution rate (K_s) value for gene pairs from Lamiales on a basis of each 1 : 1 ortholog from Dataset 1, by using ParaAT (v.2.0) (Zhang *et al.*, 2012) and KaKs_Calculator v.2.0 (Wang *et al.*, 2010). We then drew the K_s distribution and labeled the K_s peak using R/ggplot2, excluding the $K_s > 3$. Likewise, we used the 1 : 1 orthologs from Dataset 2 to label the K_s peak between Lamiales and Solanales, omitting the orthologs with $K_s > 5$.

Identification and inference of the phylogenetic location of whole gene duplication across Lamiales

In order to identify, locate and determine the WGDs in *P. huaijiensis* and other Lamiales species, we utilized a multi-pronged pipeline, including the distribution of K_s among paralogs for each species, phylogenetic reconciliation and simulation, and microsynteny among specific species.

For the K_s-based method, we applied the software 'wgd' (Zwaenepoel & Van de Peer, 2019) to construct K_s distribution (ranging from 0.05 to 3) among paralogs from eight Lamiales genomes (*T. grandis*, *H. impetiginosus*, *S. indicum*, *A. majus*, *P. huaijiensis*, *D. hygrometricum*, *O. europaea* and *F. excelsior*), and eight Didymocarpinae transcriptomes. Especially for eight Lamiales genomes, we pruned the paralogs on the basis of co-linearity analysis using i-ADHoRe (Proost *et al.*, 2012). Then, according to a fitted mixture model (BGMM in wgd), we fitted the K_s distribution of paralogs from each hypothesized WGD peak, obtained an estimation for the mean and variance of each WGD peak, and isolated those paralogs belonging to each WGD with 95% probability. The final K_s regions of two potential WGDs in *P. huaijiensis* are overlapped with corresponding K_s regions inferred from the hypothesized WGDs before and after co-linearity analysis.

For the phylogenetic approach, we used the MultiTAxon Paleopolyploidy Search (MAPS; Li *et al.*, 2015) (https://bitbucket.org/barkerlab/maps/src/master/) to locate the phylogenetic placements of the putative ancient WGDs. The MAPS algorithm works best with simple, ladderized species trees. Based on the hypothesized WGD peaks in *P. huaijiensis* and phylogeny across asterids, we clustered the orthogroups, inferred homolog trees from two datasets (Datasets 3 and 4) following the pipeline as mentioned above. Then we mapped the homolog trees of Dataset 3 (proteins from *P. huaijiensis*, *P. swinglei*, *P. fangianus*, *H. subcapitata*, *C. dispar*, *H. anachpreta* and *S. indicum*) to species tree ((((((*P. huaijiensis*, *P. swinglei*), *P. fangianus*), *H. subcapitata*), *C. dispar*), *H. anachpreta*), *S. indicum*) using MAPS tools with the parameters 'mb' (minimum bootstrap value) equal to 80 and 'mt' (the minimum percentage of the ingroup taxa to be present in all subtrees) equal to 50, to calculate the percentage of subtrees with gene duplications shared by all taxa descended from that node. Meanwhile, we calculate the percentage from 100 replicates of 1000 simulated gene trees with and without WGDs, setting the parameter 'wgd_retention_rate' as 0.20 in positive simulations. Further, we compared the percentage difference between empirical and simulated data to finally verify placements of 'younger' WGD in *P. huaijiensis*. Likewise, we mapped the homolog trees of Dataset 4 (proteins from *T. grandis*, *S. indicum*, *A. majus*, *P. huaijiensis*, *O. europaea*, *S. lycopersicum* and *V. vinifera*) to the gene tree ((((((*T. grandis*, *S. indicum*), *A. majus*), *P. huaijiensis*), *O. europaea*), *S. lycopersicum*), *V. vinifera*), to evaluate location of 'older' WGD of *P. huaijiensis*.

For the 4DTv (transversion substitutions at four-fold degenerate sites) method, we called the collinear blocks by using MCScanX (http://chibba.pgml.uga.edu/mcscan2/) with a match size of 10. Further, we calculated 4DTv values for gene pairs by using ParaAT (v.2.0) (Zhang *et al.*, 2012) and Sun's scripts (Figshare doi: 10.6084/m9.figshare.10185056). We then drew the 4DTv distribution and labeled the peak using R/ggplot2, excluding values > 0.5. For the syntenic method, we constructed and show the typical case of microsynteny between grape and sesame, and the microsynteny between grape and *P. huaijiensis*, by using MCScan (Python version; https://github.com/tanghaibao/jcvi/wiki/). In addition, we obtained and displayed the syntenic relationship of self-comparison of the *P. huaijiensis* genome by using MCScan, Circos (Darzentas, 2010) and Wgd (Zwaenepoel & Van de Peer, 2019).

Gene ontology enrichment analysis

We applied R/topGO, following the package's instructions (http://bioconductor.uib.no/2.7/bioc/vignettes/topGO/inst/doc/topGO.pdf), to analyze the gene ontology (GO) enrichment (Category: 'Molecular Function') of specific groups of genes (e.g. tandem duplications, WGDs and expanded genes), setting all *P. huaijiensis* genes as background. To avoid relatively broad annotation, here we focused only on the lowest-level GO terms under enrichment ($P < 0.01$), whereas the *P*-value was calculated using a 'classic' algorithm with Fisher's test. The lowest-level GO terms was based on the directed acyclic graph (DAG) of GO, with the parameter 'nodeSize = 100'.

Identification and comparison analysis of transcription factors in 34 eudicots

We identified types of transcription factors (TFs) among 34 typical eudicots covering the most plant family with public high-quality genome data using iTAK (Zheng *et al.*, 2016), and then classified them into detailed categories according to the PlnTFDB website (http://plntfdb.bio.uni-potsdam.de/v3.0/) (Perez-Rodriguez *et al.*, 2009). The low-frequency categories (the average number of members among 34 eudicots < 10) were excluded.

We evaluated the ranking of individual TF category (the proportion for each category out of the total genes) of *P. huaijiensis* in eudicots, according to the proportion of 34 eudicots as follows.

First, we examined whether the proportions for each category fit a normal distribution on a basis of empirical data from 34 eudicots (regarded as random sampling) using tests of normal distribution by the Kolmogorov–Smirnov method of SPSS. Then, we calculated the *z*-score (the index measure how many SDs from the mean) for each category following the format: $z\text{-score} = (x-\mu)/\sigma$, where x, μ and σ represent the proportion for individual TF category in *P. huaijiensis*, the mean proportion of 34 eudicots, and the standard deviation of 34 eudicots, respectively. Then we converted the *z*-score into normal probability, in order to evaluate the degree of proportion of individual category of *P. huaijiensis* in eudicots.

Data availability

The genome assembly of *P. huaijiensis* and sequencing data have been deposited at GenBank under Bio Project PRJNA532462. The alignments, best substitution model, phylogeny, MAPS results and all scripts are available at Figshare (doi: 10.6084/m9.figshare.10185056 and doi: 10.6084/m9.figshare.11955318).

Results

Genome assembly and characterization

We sequenced the highly heterozygous (*c.* 1%) *P. huaijiensis* genome using the Illumina next-generation sequencing platform with a series of libraries having inserts ranging from 230 bp to 15 kbp. This sequencing generated *c.* 158 Gbp clean data, yielding over 300-fold sequence depth (Table S1). The highly contiguous haploid genome assembly is 478 Mbp (Tables 1, S2), accounting for 93.5% of the estimated genome size (511 Mbp; Fig. S1). With the aid of Hi-C (*in vitro* fixation of chromosomes) technology (113 Gbp clean data, ~220 × coverage; Table S1), we anchored mounts of scaffolds into 18 pseudomolecules (Figs 1a, S2), which improved scaffold N50 to 23.5 Mbp, the largest scaffolds being 32.7 Mb (Tables 1, S2). We demonstrated a high consistency and completeness of the assembly by the mapping of 98.5% paired-end reads, 97.2% of ultra-conserved CEG, 96.5% of BUSCO, 94.5% of expressed sequence tag (EST) and 89.0–93.2% of various RNA-Seq datasets generated from different tissues and developmental stages (Tables 1, S3–S7; Fig. S3).

Table 1 Summary of genome assembly and annotation for *Primulina huaijiensis*.

	Number (percentage)
Assembly feature	
Genome-sequencing depth (×)	530
Estimated genome size (Mbp)	511
Total length of scaffolds (Mbp)	478
N50 of scaffolds (bp)	23 479 473
Total Length of contigs (Mbp)	466
N50 of contigs (bp)	28 983
Mapping rate by reads from short-insert libraries	98.5%
Core Eukaryotic Genes Mapping Approach (CEGMA) evaluation	97.2%
Benchmarking Universal Single-Copy Orthologs (BUSCO) evaluation	96.5%
EST evaluation	94.5%
RNA-Seq evaluation	89.0–93.2%
Genome annotation	
Percentage of transposable elements (TE)	54.1%
Percentage of long terminal repeat-retrotransposons (LTR)	48.4%
No. of predicted protein-coding genes	31 328
No. of genes annotated to public database	30 583 (97.6%)
No. of genes annotated to GO database	18 781 (59.9%)
No. of genes duplicated by tandem duplications	1948 (6.2%)
No. of genes duplicated by syntenic duplications	15 197 (48.5%)
No. of genes duplicated by the *D* event	10 132 (32.3%)

GO, gene ontology.

We found that 54.1% of the assembly is covered with transposable elements (TEs), mostly long terminal repeat-retrotransposons (LTR; Fig. 1b), making up 48.4% of the genome (Tables 1, S8). Using a variety of gene-modeling software and databases for gene annotation, we identified a total of 31 328 protein-coding genes (Fig. 1c; Table S9). Of these genes, homologs of 97.5% were identified in public protein databases (Tables 1, S10). Tandem duplicates (Fig. 1d) occurred for 6.2% of the genes and were preferentially enriched in transferase activity (Fig. S4).

A comparison of the predicted proteomes of *P. huaijiensis* and 16 other sequenced angiosperms indicated that 5292, 7114 and 10 791 orthogroups were shared between *P. huaijiensis* and angiosperms, Lamiales and Gesneriaceae, respectively. Moreover, we identified 38 genes from 11 orthogroups and 2322 single-copy genes that were specific to *P. huaijiensis* (Fig. 2).

Primulina huaijiensis experienced two rounds of WGD after the γ event

We utilized a combination of K_s (synonymous substitution rate)-based (WGD; Zwaenepoel & Van de Peer, 2019), phylogenetic (MAPS; Li *et al.*, 2015) and syntenic (MCSCAN, https://github.com/tanghaibao/jcvi/wiki/) approaches to identify at least five WGDs across Lamiales phylogeny, including 16 Lamiales taxa derived from eight genomes and eight Didymocarpinae transcriptomes (Table S11), plus three outgroups (tomato, potato and grape) (Fig. 3a). Analysis of the duplicates from *P. huaijiensis* genomes revealed three K_s peaks, which are indicative of three WGDs, herein named as *D* (K_s range: 0.050–0.302), *L* (K_s range: 0.640–1.407) and γ (shared by all the eudicots), respectively (Fig. 3b).

The *D* event is a novel lineage-specific WGD event. Combined with K_s distribution for the family Gesneriaceae (Figs 3b,c, S5) and the phylogeny (Fig. 3a; Roalson & Roberts, 2016), we inferred that the *D* event was shared by almost all of the subtribe Difymocarpinae, excluding the *Henckelia* genus. The placement of a lineage-specific *D* event was further supported with the MAPS analysis (Fig. 3d) and the variation of chromosome numbers in 10 species from family Gesneriaceae (Fig. 3a). According to the species divergence time in Gesneriaceae, the age of the *D* event was estimated at *c.* 20.6–24.2 Ma (Fig. 3a), slightly earlier than the mid-Miocene Climatic Optimum (16–18 Ma).

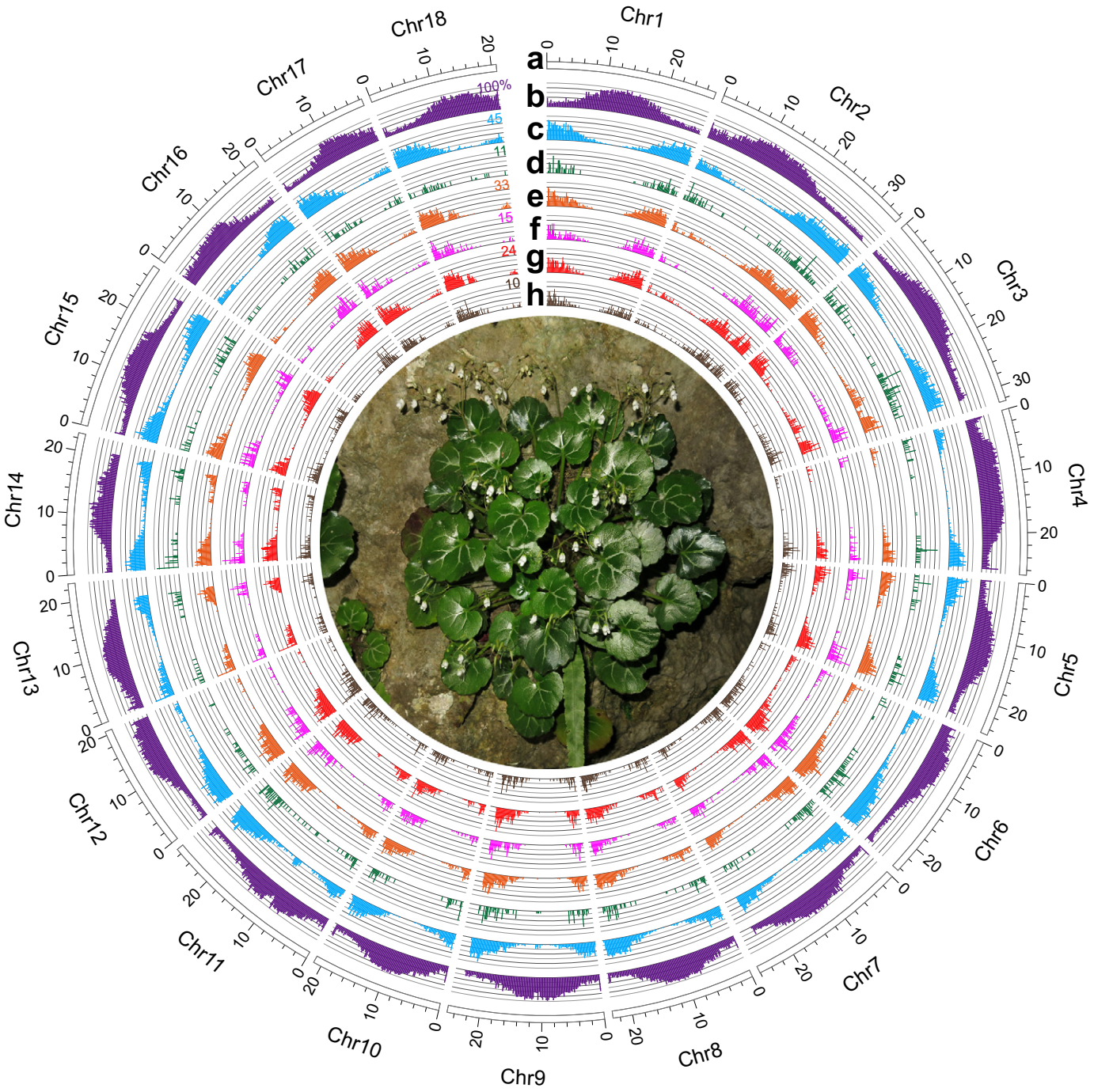


Fig. 1 *Primulina huaijiensis* genome and photo. Landscape of the *P. huaijiensis* genome, comprising 18 pseudomolecules that cover ~93.5% of assembly (a); concentric circles, from outermost to innermost, showing transposable element (TE) percentage (purple; b); gene density (blue; c); density of duplicates resulted from tandem duplications (green; d); density of duplicates located in all syntenic duplicates (orange; e); density of duplicates from the *L* event (pink; f), density of duplicates from the *D* event (red; g), density of transcription factors (TFs) (brown; h) in each 200-kbp nonoverlapping window.

The *L* event corresponds to known WGD events in other species (Edger *et al.*, 2017; Sollars *et al.*, 2017; Unver *et al.*, 2017; Ren *et al.*, 2018; Li & Barker, 2020). For instance, Sollars *et al.* (2017) found a K_s peak shared by *F. excelsior* and monkey flower (belongs to the lineage of *S. indina* and *T. grandis*), but this was not supported by synteny analysis. Edger *et al.* (2017) found support for two possible *L* events: one is an order-wide WGD event and the other occurred after the divergence of Oleaceae. Nevertheless, they stated that it was unclear if this was two events or just one with skewed signal. Unver *et al.* (2017) and Li & Barker (2020) recognized two lineage-specific WGDs for *O. europaea* and *F. excelsior*, whereas the *L* event is shared by other Lamiales. We recovered 7524 gene trees of homologs that were used for mapping gene duplication events to the species tree surrounding the *L* event. This result supports the hypothesis that the *L* event is shared by almost all the Lamiales, excluding the

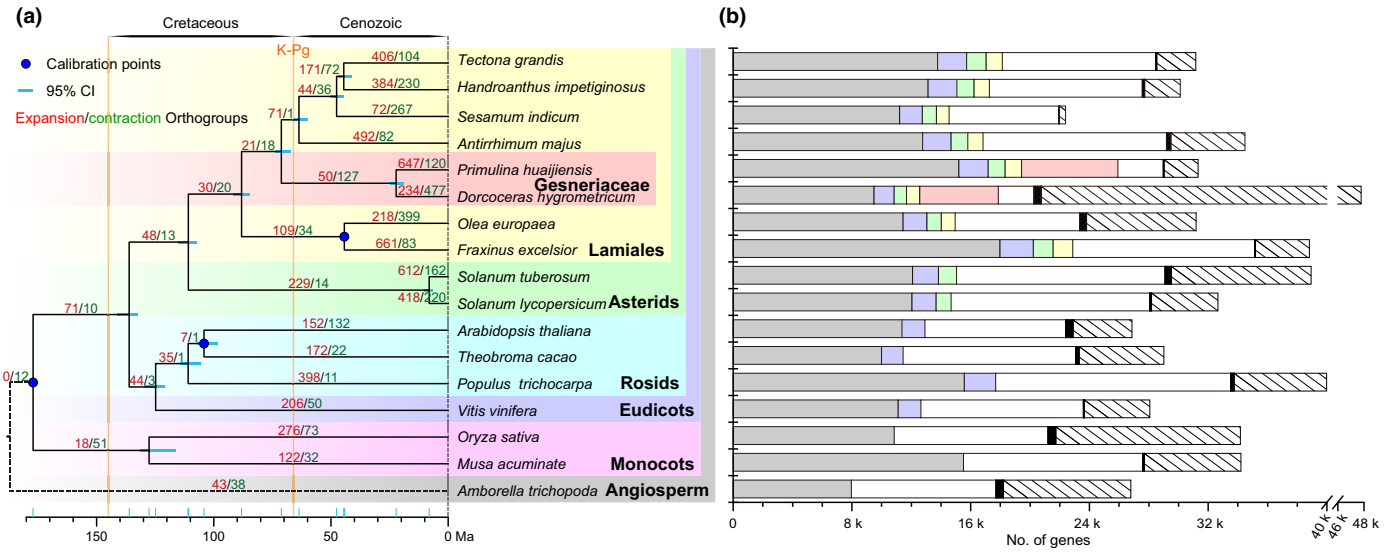


Fig. 2 Evolution of *Primulina huaijiensis* genome and orthogroups. (a) The phylogeny, divergence time and orthogroup expansions/contractions for 19 angiosperms. The tree was constructed by maximum-likelihood (ML) method using 583 single copy orthogroups. All nodes have 100% bootstrap support. Divergence time was estimated on a basis of three calibration points (blue circles). CI, confidence interval. The numbers in red and green indicate the numbers of orthogroups that have expanded and contracted, respectively, along particular branches. K-Pg, Cretaceous–Palaeogene period. (b) The comparison of genes among 19 angiosperms. Gray bars, genes belonging to 5292 angiosperm-shared orthogroups in each of 19 angiosperms; gray + blue bars, genes belonging to 6015 eudicot-shared orthogroups in each of 16 eudicots; gray + blue + green bars, genes belonging to 6496 asterid-shared orthogroups in each of ten asterids; grey + blue + green + yellow bars, genes belonging to 7114 asterid-shared orthogroups in each of eight Lamiales plants; grey + blue + green + yellow + pink bars, genes belonging to 10 791 Gesneriaceae-shared orthogroups in each of two Gesneriaceae species; striped and black bars, genes belonging to species-specific single-copy genes and orthogroups, respectively; white bars, the remaining genes for each genome.

lineages of the family Oleaceae (Fig. 3e). The location of the *L* event also coincided with the result of 4DTv distribution (Fig. 3f).

Visualization of microsynteny shows that there are approximately two copies of each syntenic block from grape in sesame, and four copies in *P. huaijiensis* (Fig. 3g), indicating sesame and *P. huaijiensis* had one and two rounds of WGD since the γ event, respectively, which is consistent with the K_s and phylogenetic result.

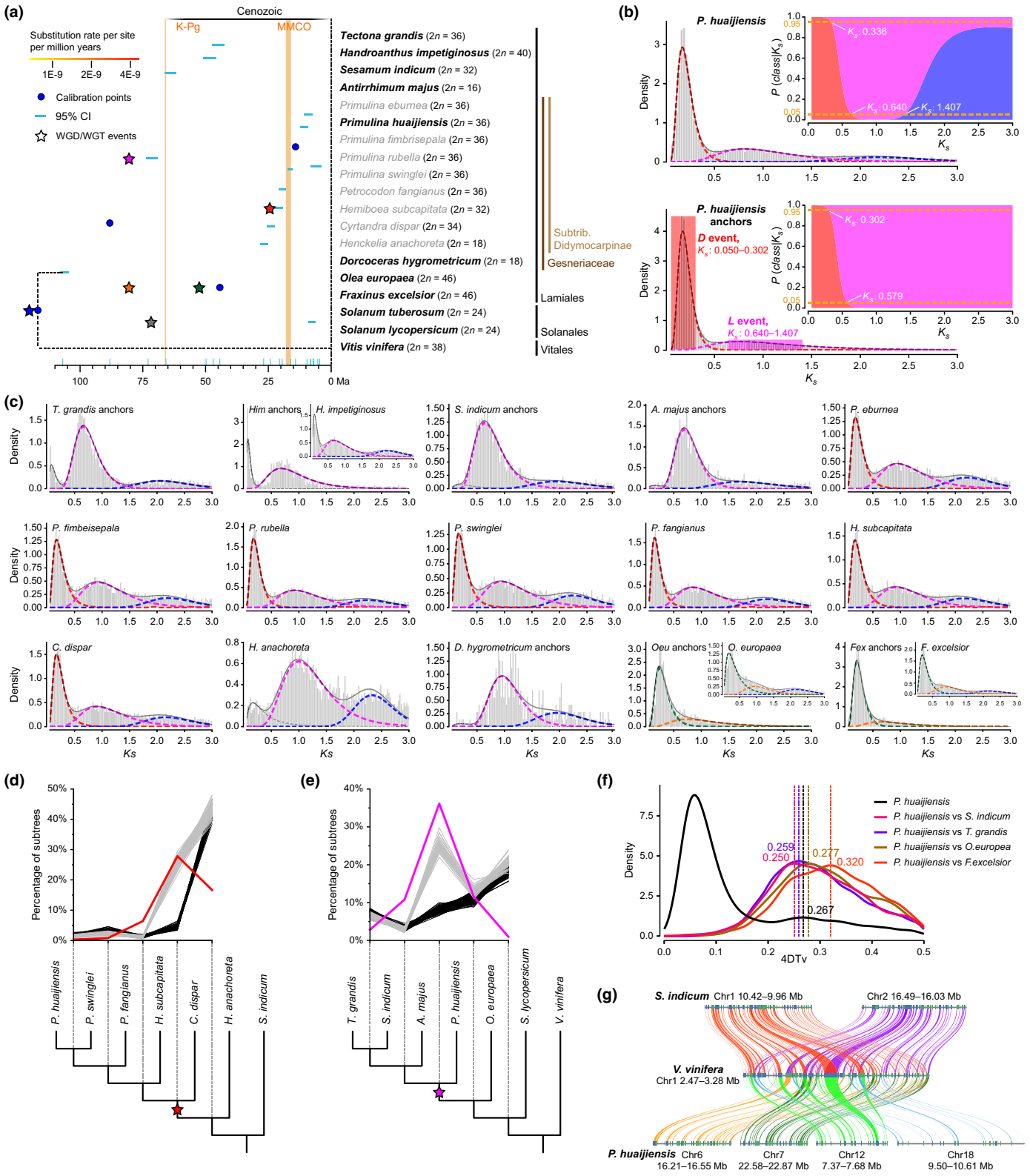
Differential retention of duplicates among different WGDs

Following syntenic duplications (WGDs and segmental duplication events), some gene duplicates are eliminated or inactivated, and thus return to a single-copy state, whereas others are retained, and these surviving duplicates can contribute to physiological innovations and evolutionary adaptation (Li *et al.*, 2016). We found that syntenic duplicates occurred for 48.5% of the genes, and the proportions of retained duplicates differed among the two WGD events in *Primulina*, with 10 132 (32.3% of the total genes) and 4123 (13.2%) of duplicates being retained from the *D* and *L* events, respectively (Figs 1e–g, 3b, S6).

In order to gain insight into the functions of retained genes following the two individual WGDs, we determined whether there was enrichment of specific molecular function from GO categories for genes with K_s values in the regions associated with the two WGDs (Figs S7–S9). We found that three GO terms ('sequence-specific DNA binding', 'chromatin binding' and 'uniquitin-protein transferase activity') were over-represented in retained genes of both WGD events (Table 2; Figs S8, S9). Specifically, GO terms related to 'protein binding' and 'zinc ion binding' were significantly enriched ($P < 1E\text{-}5$) for the *D* event, but were not significantly enriched for the *L* WGDs. Over-representation of these categories exhibits a similar pattern for 'protein tyrosine kinase activity' (related to PKs, protein kinases) and 'secondary active transmembrane transport' (Table 2). For the *L* event, a number of GO categories exhibited enrichment that was not apparent in the *D* event: 'DNA binding transcription factor activity', 'structural constituent of ribosome', 'molecular function regulator' and 'ligase activity' (Table 2). These results suggested not only that different proportions of duplicates are retained after different WGDs, but also that different functional categories are retained in different WGDs.

Gene family expansion is due mainly to lineage-specific WGD

In angiosperms, gene family expansion is an important facilitator of evolutionary adaptation and trait innovations (Ohno, 1970; Wang *et al.*, 2019). Expansion can occur via different rounds of WGDs and/or tandem duplications. We attempted to ascertain the relative contribution of these different processes to gene family expansion in *Primulina*. Using CAFE software (De Bie *et al.*, 2006), we estimated the ancestral gene content at each node of the species tree covering 17 taxa across the angiosperm, and modeled the significant changes along each branch (Fig. 2). This analysis indicated that *P. huaijiensis* has 647 expanded orthologous gene families, containing 4038 genes, compared to the inferred ancestral



Gesneriaceae genome. Of the 4038 expanded genes in *P. huaijiensis*, 54.7% resulted from all syntenic duplications, significantly higher than the percentages for all genes, 48.5% (P = 1.3E-6) (Table 3). In particular, 1410 genes (34.9%) in expanded gene families were retained following the *D* event, much larger than the number following the *L* event (980 genes, 24.3%), indicating that the lineage-specific WGD (*D* event) contributed greatly to gene family expansion (Table 3).

Fig. 3 The identification and phylogenetic location of whole genome duplication (WGD) and whole genome triplication (WGT) events in *Primulina huaijiensis* and other Lamiales species. (a) The phylogenetic tree across asterids shows the topology, divergence time, substitution rate and WGD/WGT events. The tree was constructed by a maximum-likelihood (ML) method using 482 single copy orthogroups. Species with genomes are in black and in bold, species without genome are in gray. The numbers in parentheses near each species indicate the number of chromosomes in corresponding species. All nodes have 100% bootstrap support. Divergence time was estimated on a basis of four calibration points (blue circles). CI, confidence interval. The color along each branch of the phylogeny shows the variations in the substitution rate of all sites (yellow, slow; red, fast). The putative WGD/WGT events are depicted by stars. K-Pg and MMCO are abbreviations for the Cretaceous–Palaeogene period and Mid-Miocene Climatic Optimum period, respectively. (b) K_s distribution for paralogs in *P. huaijiensis*, before (upper part) and after (bottom part) co-linearity analysis. Dashed lines, representing duplicates from individual WGDs, are fitted by a fitted mixture model (BGMM). The two graphs in the upper-right corners indicate the proportion of paralogous pairs from hypothesized WGDs at different K_s values. Gray columns in red and pink background indicate the duplicates from the *D* and *L* events in *P. huaijiensis*, respectively. The final K_s regions of two potential WGDs in *P. huaijiensis* are overlapped with corresponding K_s regions inferred from the hypothesized WGDs before and after co-linearity analysis. (c) K_s distribution for paralogs in 15 Lamiales species. The species with or without genomes show the K_s distribution of paralogs from corresponding species after or before co-linearity analysis, respectively. For some plants with genomes, the results for before the co-linearity analysis are shown as inserts because duplicates from γ event would be hidden by the co-linearity analysis. (d, e) MultitAxon Paleopolyploidy Search (MAPS) results on the portion of the phylogeny surrounding potential WGDs. Percentage of subtrees indicates percentage of duplicates shared by descendant species at each node; results are portrayed for observed data (red line and pink line), 100 repetitions of null simulations (black lines) and positive simulations (gray lines). The red and pink stars represent the *D* and *L* events in *P. huaijiensis*, respectively. (f) Four-fold degenerate (4DTv) distributions with HKY substitution models for *P. huaijiensis*, *P. huaijiensis* vs *Sesamum indicum*, *P. huaijiensis* vs *Tectona grandis*, *P. huaijiensis* vs *Fraxinus excelsior*, and *P. huaijiensis* vs *Olea europaea*. (g) Microsynteny among grape, sesame and *P. huaijiensis*.

Table 2 Shared and differential retention between two whole genome duplication (WGD) events in the *Primulina huaijiensis* genome.

			The *D* event		The *L* event	
GO ID	Term annotated	No. of background	No. of genes	*P*-value[a]	No. of genes	*P*-value
GO:0043565	Sequence-specific DNA binding	542	244	**2.40E-07**	125	**1.70E-09**
GO:0003682	Chromatin binding	418	169	**0.0072**	86	**6.40E-05**
GO:0003700	DNA binding transcription factor activity	823	301	0.1226	200	**5.30E-17**
GO:0004713	Protein tyrosine kinase activity	1186	452	**0.0051**	190	0.0111
GO:0008270	Zinc ion binding	1285	525	**8.00E-07**	205	0.0101
GO:0015291	Secondary active transmembrane transporter activity	121	55	**0.0087**	15	0.7056
GO:0004842	Ubiquitin-protein transferase activity	117	54	**0.0064**	27	**0.0042**
GO:0005515	Protein binding	3991	1549	**1.40E-10**	572	0.1112
GO:0016616	Oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	162	72	**0.0059**	18	0.8639
GO:0003735	Structural constituent of ribosome	443	155	0.4544	82	**0.0027**
GO:0098772	Molecular function regulator	228	79	0.5229	46	**0.0044**
GO:0016879	Ligase activity, forming carbon–nitrogen bonds	141	58	0.6295	30	**0.0090**

[a]The *P* values < 0.01 are in bold.

Applying GO enrichment analysis to the expanded genes revealed that there were at least eight significantly enriched GO terms belonging to category 'Molecular Function' (Table 3; Fig. S10), which could be classed into three major categories: (1) TFs (GO:0003700, GO:0043565, GO:0003682), which are extremely significantly expanded in the phylogeny, especially the first two terms with $P < 1E\text{-}30$; (2) ion binding and transport (the terms of 'cation-transporting ATPase activity', 'zinc ion binding' and 'calcium ion binding'); and (iii) others, such as PKs (GO:0004713), which were more likely to be retained following the *D* event (41.8%), than following the *L* event (24.2%) (Table 3).

Preferential retention of TFs following WGDs

Eudicot TFs belong to superfamilies with hundreds to thousands of copies, and duplicated copies may have important roles in adaptive evolution (Lehti-shiu & Shiu, 2012). Duplication of genes by WGD has the potential to free one of the copies to evolve novel functions (Ren *et al.*, 2018), and thus potentially provides a major source of raw material for adaptation to novel environments. To confirm the expansion of TFs in *P. huaijiensis*, and to identify the specific category that may have contributed to adaptation to karst environments, we classified and compared individual TFs from 34 eudicots that represent the most plant families with public high-quality genome data, and further examined whether these genes were retained from WGDs or tandem duplications.

We identified 2536 TFs in the *P. huaijiensis* genome, occurring for 8.1% of the total genes (Fig. 1h; Table S12). The proportion that were TFs ranked the second highest in *P. huaijiensis*, after *Actinidia chinensis*, among the 34 eudicots, which may be due to their extremely significant

Table 3 Gene ontology (GO; molecular function) enrichment analysis of gene family expansions in *Primulina huaijiensis*.

GO ID	Term annotated	No. of background	No. of expanded genes	*P*-value	TF%[a]	*D*%[b]	*L*%[c]	WGD%[d]	TD%[e]
GO:0003700	DNA binding transcription factor activity	823	270	< 1E-30	97.0	35.9	32.6	59.6	9.3
GO:0043565	Sequence-specific DNA binding	542	173	< 1E-30	95.4	44.5	29.5	67.1	4.0
GO:0003682	Chromatin binding	418	106	6.50E-12	96.2	30.2	19.8	57.5	13.2
GO:0019829	Cation-transporting ATPase activity	120	44	5.00E-11	0	0	0	0	4.5
GO:0008270	Zinc ion binding	1285	211	2.90E-08	14.2	43.6	33.6	69.7	7.6
GO:0005509	Calcium ion binding	271	57	2.30E-06	0	15.8	0	15.8	3.5
GO:0004713	Protein tyrosine kinase activity	1186	194	0.00048	0.5	41.8	24.2	54.1	5.7
GO:0003924	GTPase activity	203	41	0.00294	0	22.0	9.8	26.8	7.3
Total (*P*-value)[f]		31 328	4038	–	16.9 (< 1E-30)	34.9 (0.00783)	24.3 (< 1E-30)	54.7 (1.3E-6)	9.5 (4.3E-18)

[a]The percentage of transcription factors (TF) in categories of expanded genes with specific GO terms; [b]The contribution rate by the *D*-WGD event in each GO term and sum of expansion genes; [c]the contribution rate by the *L*-WGD event in each GO term and sum of expansion genes; [d]the contribution rate by all WGD events (including small-scale segmental duplications) in each GO term and sum of expansion genes; [e]the contribution rate by tandem duplications (TD) in each GO term and sum of expansion genes; [f]*P*-value for the enrichment of genes related to TF, the *D* event, *L* event, WGD and TD in expanded genes, compared to that in total *P. huaijiensis* genes.

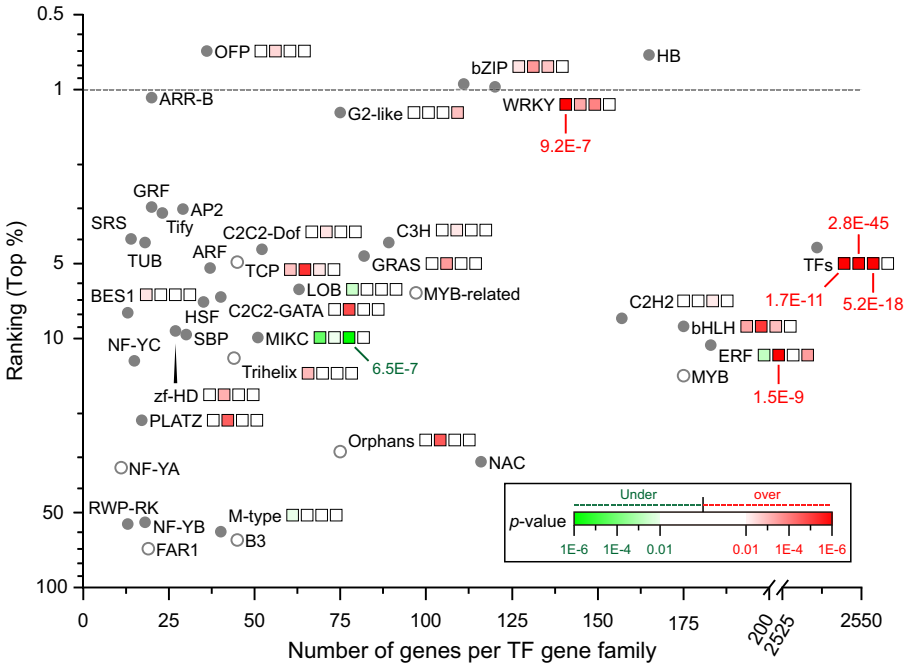


Fig. 4 Analysis of transcription factors (TF) in *Primulina huaijiensis*. The *x*-axis shows the number of individual TFs in *P. huaijiensis*, whereas the *y*-axis indicates the ranking of corresponding category – how *P. huaijiensis* ranks, compared to other eudicots, with respect to the extent to which a gene family has expanded. The solid and hollow circles indicate whether the proportion of individual TF category (out of the total genes) of 34 eudicots fits a normal distribution or not, respectively. The number of genes in each category in the 34 eudicots is listed in Supporting Information Table S12. The squares from left to right represent the significance of over- or under-retention of duplicates from the *D* event, *L* event, syntenic duplications and tandem duplications, respectively, with color intensity indicating the corresponding *P*-value. Categories with *P*-value > 0.01 are omitted, whereas the ones which have extremely significantly retention ($P < 1E\text{-}6$) from specific duplications were shown in details around the corresponding square.

retention following both of two WGDs (*D* and *L* events; Fig. 4; Table S12). Compared to other eudicots, three TF categories (WRKYs, HBs and bZIPs) are over-represented in *P. huaijiensis*. Each of these categories has ≥100 copies, which ranks in the top 1% among all eudicots (Fig. 4b). Remarkably, among all the TF categories, the duplicates of WRKYs are most preferentially retained following the *D* event ($P < 1E\text{-}6$), as well as following all of the syntenic duplicates. Additionally, the duplicates of bZIPs are more likely to be retained from both of the *D* and *L* events.

Discussion

Lineage-specific WGD probably linked with speciation diversification

The prevalence of whole genome duplication (WGD) in angiosperms has long been acknowledged (Wood *et al.*, 2009; Jiao *et al.*, 2011), yet the impact of WGD events on species diversification is a subject of debate (Mayrose *et al.*, 2011; Soltis *et al.*, 2014). Previous analyses of the impact of WGDs on diversification have generally focused on highly diverged clades representing deep divergences at the tribe level, family level or above (Estep *et al.*, 2014; Tank *et al.*, 2015; Landis *et al.*, 2018). By contrast, few studies have examined the relationship between rates of species diversification and WGDs for shallower divergences (Clarkson *et al.*, 2017). Our analyses revealed that in addition to a WGD (the *L* event) that is shared with almost all Lamiales, excluding lineages in the family Oleaceae, the subtribe Didymocarpinae experienced a lineage-specific WGD (*D* event) occurring *c.* 20.6–24.2 Myr ago (Ma) (Fig. 3a). Kong *et al.* (2017) identified an early burst of speciation in the *Primulina* genus at *c.* 14.14 Ma. This time period is 6.5–10 Myr later than the *D* event (Fig. 3a) (Kong *et al.*, 2017), which is consistent with the 'lag-time model' (Schranz *et al.*, 2012), where increase in diversification rates tend to follow WGD events after a lag time of millions of years (Schranz *et al.*, 2012; Tank *et al.*, 2015; Landis *et al.*, 2018).

The Gesneriaceae is a mid-sized to large family comprising approximately 3300 species in 160 genera belonging to three subfamilies: Sanangoideae (monospecific genus *Sanango* in Andes), Gesnerioideae (New World) and Didymocarpoideae (Old World) (Weber *et al.*, 2013). The subfamily Didymocarpoideae was further divided into two tribes, with each consisted of five subtribes. Of them, Didymocarpinae is the largest subtribe of approx. 30 genera and ≤ 1600 species (Weber *et al.*, 2013). There are seven big genera containing > 100 species in the subfamily Didymocarpoideae, four of which, including *Crytandra* (the largest genus with approx. 800 species), *Primulina* (approx. 180 species), *Aeschynanthus* (approx. 160 species) and *Oreocharis* (*c.* 140 species), belong to Didymocarpinae, and experienced a shared *D* event. Roalson & Roberts (2016) identified elevated diversification rates in several lineages in Didymocarpinae, including *Cyrtandra*, *Oreocharis*, *Hemiboea* and *Primulina*. Hence, the lineage-specific WGD (*D* event) is likely to have played an important role in species diversification in the subtribe Didymocarpinae.

Differential retention after WGDs

Despite the repeated occurrence of WGDs across angiosperms, gene number and genome size do not remain doubled after each event because of subsequent fractionation processes (Jiao, 2018). Following a WGD, rapid and large-scale duplicate loss typically occurs within a few Myr. Several empirical studies have demonstrated that the proportion of duplicates retained over time usually decays more-or-less exponentially (Li *et al.*, 2016; Ren *et al.*, 2018). Whether the loss of gene duplicates is a random or nonrandom process is still debated. Ren *et al.* (2018) found a constant stochastic loss of gene duplicates, especially for 'younger' recent WGDs. However, Li *et al.* (2016) observed that gene retention following WGDs exhibited a highly nonrandom pattern, with a fraction of duplicates often being retained for long periods, or even indefinitely (Lynch & Conery, 2000; Maere *et al.*, 2005; Li *et al.*, 2016). *Primulina huaijiensis* experienced two rounds of WGD after the γ event, providing an excellent opportunity to investigate differential retention of duplicates following WGDs. We found differential function enrichment of retained genes following these two individual WGDs, revealing a deviation from random decay in this species.

A previous study in *Arabidopsis thaliana* revealed that duplicates of some gene families that were retained after one WGD also are preferentially retained after a subsequent WGD (Seoighe & Gehring, 2004). In agreement with this finding, two gene ontology (GO) terms related to transcription factors (TFs) ('sequence-specific DNA binding' and 'chromatin binding') and the GO term 'uniquitin-protein transferase activity' are retained following both of the *D* and *L* events in *P. huaijiensis* (Table 2). However, some gene groups do not obey this pattern, such as the genes encoding 'DNA binding transcription activity' (this term contains 183 ERFs (ethylene response factors), 51 MIKCs (the transcription factors containing four conserved domains: the MADS-box (M-) domain, the intervening (I-) domain, the keratin-like (K-) domain, and the C-terminal (C-) domain), and 551 other TFs) retained following the *L* event with *P*-value < 0.01, but not retained after the *D* event (Table 2). Although in general TFs were retained following both of the *L* and *D* events, different TF categories have independent patterns. For example, ERFs and MIKCs were significantly under-represented following the *D* event (Fig. 4). A similar pattern also was exhibited by the GO term 'structural constituent of ribosome'. Preferential gene retention has been widely suggested to be associated with key phenotypic novelty and adaptation to environmental changes (Hegarty & Hiscock, 2008; Soltis & Soltis, 2016). Duplicate copies resulting from 'older' WGDs are more likely to have undergone neofunctionalization or subfunctionalization, whereas duplicates produced by the most recent WGDs also might be retained because they increase gene dosage. Ren *et al.* (2018) found that recent WGDs allow ancestral duplicates to be lost, presumably because the new additional copies reduce purifying selection on older duplicates, thereby accelerating their rate of loss. We found that the duplicates related to protein tyrosine kinase activity were significantly ($P < 0.01$) and largely retained from the *D* event (Fig. 3b), but not the *L* event, suggesting that new duplications in this gene family may render older duplications more expendable. Previous findings that expansions of protein kinases have played important roles in adaptive evolution (Lehti-Shiu & Shiu, 2012), provide clues for seeking the key duplicates that enhance adaptation by plants to harsh environments.

Gene expansions associated with habitat adaptation

Understanding the mechanisms through which genome duplication can result in evolutionary novelty remains a challenge. One of

the obvious consequences of WGDs is the simultaneous creation of gene duplicates of the whole genome, which have long been thought to constitute a major source of new material for adaptation (Ohno, 1970). Recently, Wu *et al.* (2020) investigated the survivors of gene duplicates in 25 selected genomes, and found retained duplicates following WGDs have functions in adaption to dramatic environmental changes, for example, retentions following WGDs around the K-Pg (Cretaceous–Palaeocene) boundary were commonly enriched for the genes in response to low temperature and darkness. Our comparative genomic analysis revealed that *P. huaijiensis* has experienced expansions of many gene families, and that much of these expansion can be ascribed to the WGDs (Table 3). The GO analysis of expansions showed that the terms related to TFs were significantly more enriched in *P. huaijiensis* (Table 3), similar to previous studies (Maere *et al.*, 2005; Wu *et al.*, 2020). In particular, WRKYs, the 6th largest-size family of TFs in *P. huaijiensis*, were the most preferentially maintained from all syntenic duplicates, as well as following the *D* event (Fig. 4). By contrast, Wu *et al.* (2020) examined the retention pattern of duplicates of 59 TF categories following recent waves of four independent WGDs in *Tarenaya hassleriana*, *Glycine max*, *Panicum virgatum* and *Zea mays*, and uncovered retention of WRKY below the average ranking among TFs, only listed 32nd, 51st, 42nd and 29th, respectively. In addition, *P. huaijiensis* was found herein to rank in the top 1% of species in the proportion of WRKYs among eudicots (Fig. 4; Table S12). This result suggests to us that there is something special about WRKYs being retained after WGD *D*, and that expansions of WRKY may have played a key role in evolutionary adaptation or trait innovations. It also is known that WRKYs can interact with calmodulin (regulated by calcium ion (Ca^{2+}) fluxes), resistance proteins and other WRKYs, leading to pivotal roles in ameliorating drought- or salt-tolerance (Jiang & Deyholes, 2009; Wu *et al.*, 2009; Rushton *et al.*, 2010). The karst soil environment is characterized by high salinity (especially Ca^{2+} and magnesium ion (Mg^{2+})) and low water content. This, in turn, suggests that expansions of WRKYs thus may have facilitated adaptation by *P. huaijiensis* to karst limestone habitats. In addition, several TFs specifically related to salt and drought stress also became enriched following the *D* event; for example, bZIPs are implicated in salt/drought stress signaling (Singh *et al.*, 2002). These findings, along with the expansions of 'ion binding' and 'protein tyrosine kinase', suggest that this species has evolved a complex physiological system that allows it to survive in extreme and harsh cave environments.

Conclusions

We have produced a high-quality genome assembly of *P. huaijiensis*, a cave-dwelling plant in karst habitats. This is the first chromosome-level genome in Gesneriaceae. A combination of K_s-based, tree-based and syntenic approaches showed that *P. huaijiensis* experienced two rounds of WGD since the γ triplication event shared by all eudicots. The ancient one (the *L* event) was shared with almost all Lamiales, whereas the latest one (the *D* event) was shared by almost the entire subtribe Didymocarpinae. The *D* event occurring around a period of the early Miocene might have facilitated species diversification in Didymocarpinae. We found biased retention of duplicates for the *D* event, which have contributed to gene family expansion of genes coding for WRKYs, as well as other TFs (bZIP) and genes related to ion binding and protein tyrosine kinase. The evidence presented here suggests that the lineage-specific WGD event is likely to have made a major contribution to the adaptation of *P. huaijiensis* and potentially other *Primulina* species to the limestone karst habitats.

Acknowledgements

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (XDB31000000) and the National Natural Science Foundation of China (U1501211, 31501799). We thank Zheng Li (University of Arizona) for suggestions and discussion on WGD identification.

Author contributions

MK and MR conceived the project and designed the study; Chao Feng, LY and Chen Feng performed the sampling and experiments; Chao Feng, KW, JW, LW and HK performed the data analysis; Chao Feng designed and visualized the figures; MK, Chao Feng and MR wrote the manuscript; and all authors read and approved the final manuscript.

ORCID

Ming Kang https://orcid.org/0000-0002-4326-7210
Mark Rausher https://orcid.org/0000-0002-6541-9641

References

Ai B, Gao Y, Zhang XL, Tao JJ, Kang M, Huang HW. 2015. Comparative transcriptome resources of eleven *Primulina* species, a group of 'stone plants' from a biodiversity hot spot. *Molecular Ecology Resources* **15**: 619–632.

Burton JN, Adey A, Patwardhan RP, Qiu RL, Kitzman JO, Shendure J. 2013. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nature Biotechnology* **31**: 1119–1125.

Call VB, Dilcher DL. 1992. Investigations of angiosperms from the Eocene of southwestern North America: Samaras of *Fraxinus wilcoxiana* Berry. *Review of Palaeobotany and Palynology* **74**: 249–266.

Chung KF, Leong WC, Rubite RR, Repin R, Kiew R, Liu Y, Peng CI. 2014. Phylogenetic analyses of Begonia sect. Coelocentrum and allied limestone species of China shed light on the evolution of Sino-Vietnamese karst flora. *Botanical Studies* **55**: 1.

Clarkson JJ, Dodsworth S, Chase MW. 2017. Time-calibrated phylogenetic trees establish a lag between polyploidisation and diversification in *Nicotiana* (Solanaceae). *Plant Systematics and Evolution* **303**: 1001–1012.

Clements R, Sodhi NS, Schilthuizen M, Ng PKL. 2006. Limestone karsts of southeast Asia: Imperiled arks of biodiversity. *BioScience* **56**: 733–742.

Cui LY, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A *et al.* 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Research* **16**: 738–749.

Darzentas N. 2010. Circoletto: visualizing sequence similarity with Circos. *Bioinformatics* **26**: 2620–2621.

De Bie T, Cristianini N, Demuth JP, Hahn M. 2006. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**: 1269–1271.

Doyle JJ. 1990. Isolation of plant DNA from fresh tissue. *Focus* **12**: 13–15.

Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, Wendel JF. 2008. Evolutionary genetics of genome merger and doubling in plants. *Annual Review of Genetics* **42**: 443–461.

Edger PP, Heidel-Fischer HM, Bekaert M, Rota J, Gloeckner G, Platts AE, Heckel DG, Der JP, Wafula EK, Tang M *et al.* 2015. The butterfly plant arms-race escalated by gene and genome duplications. *Proceedings of the National Academy of Sciences, USA* **112**: 8362–8366.

Edger PP, Smith R, McKain MR, Cooley AM, Vallejo-Marin M, Yuan YW, Bewick AJ, Ji LX, Platts AE, Bowman MJ *et al.* 2017. Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. *Plant Cell* **29**: 2150–2167.

Ellegren H. 2014. Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution* **29**: 51–63.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**: 157.

Estep MC, McKain MR, Diaz DV, Zhong JS, Hodge JG, Hodkinson TR, Layton DJ, Malcomber ST, Pasquet R, Kellogg EA. 2014. Soltis & Soltis Allopolyploidy, diversification, and the Miocene grassland expansion. *Proceedings of the National Academy of Sciences, USA* **111**: 15149–15154.

Feng C, Xu MZ, Feng C, von Wettberg EJB, Kang M. 2017. The complete chloroplast genome of *Primulina* and two novel strategies for development of high polymorphic loci for population genetic and phylogenetic studies. *BMC Evolutionary Biology* **17**: 224.

Foster CSP, Sauquet H, van der Merwe M, McPherson H, Rossetto M, Ho SYW. 2017. Evaluating the impact of genomic data and priors on Bayesian estimates of the angiosperm evolutionary timescale. *Systematic Biology* **66**: 338–351.

Fu LM, Niu BF, Zhu ZW, Wu ST, Li WZ. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**: 3150–3152.

Gandolfo MA, Nixon KC, Crepet WL. 1998. A new fossil flower from the Turonian of New Jersey: *Dressiantha bicarpellata* gen. et sp. nov. (Capparales). *American Journal of Botany* **85**: 964–974.

Gao Y, Ai B, Kang M, Huang H. 2015. Geographical pattern of isolation and diversification in karst habitat islands: a case study in the *Primulina eburnea* complex. *Journal of Biogeography* **42**: 2131–2144.

Geiser C, Mandáková T, Arrigo N, Lysak MA, Parisod C. 2016. Repeated whole-genome duplication, karyotype reshuffling, and biased retention of stress-responding genes in Buckler mustard. *Plant Cell* **28**: 17–27.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng QD *et al.* 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**: 644–652.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M *et al.* 2013. *De novo* transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nature Protocols* **2013**: 1494–1512.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biology* **9**: R7.

Hao Z, Kuang YW, Kang M. 2015. Untangling the influence of phylogeny, soil and climate on leaf element concentrations in a biodiversity hotspot. *Functional Ecology* **29**: 165–176.

Hegarty MJ, Hiscock SJ. 2008. Genomic clues to the evolutionary success of review polyploid plants. *Current Biology* **18**: R435–R444.

Huang J, Zhang CM, Zhao X, Fei ZJ, Wan KK, Zhang Z, Pang XM, Yin X, Bai Y, Sun XQ *et al.* 2016. The Jujube genome provides insights into genome evolution and the domestication of sweetness/acidity taste in fruit trees. *PLoS Genetics* **12**: e1006433.

Jiang YQ, Deyholos MK. 2009. Functional characterization of Arabidopsis NaCl-inducible WRKY25 and WRKY33 transcription factors in abiotic stresses. *Plant Molecular Biology* **69**: 91–105.

Jiao YN. 2018. Double the genome, double the fun: genome duplications in angiosperms. *Molecular Plant* **11**: 357–358.

Jiao YN, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang HY, Soltis PS *et al.* 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**: 97–100.

Kang M, Tao JJ, Wang J, Ren C, Qi QW, Xiang QY, Huang HW. 2014. Adaptive and nonadaptive genome size evolution in Karst endemic flora of China. *New Phytologist* **202**: 1371–1381.

Katoh K, Standley DM. 2013. MAFFT Multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology Evolution* **30**: 772–780.

Kong HH, Condamine FL, Harris AJ, Chen JL, Pan B, Moller M, Hoang VS, Kang M. 2017. Both temperature fluctuations and East Asian monsoons have driven plant diversification in the karst ecosystems from southern China. *Molecular Ecology* **26**: 6414–6429.

Landis JB, Soltis DE, Li Z, Marx HE, Barker MS, Tank DC, Soltis PS. 2018. Impact of whole-genome duplication events on diversification rates in angiosperms. *American Journal of Botany* **105**: 348–363.

Lehti-Shiu MD, Shiu SH. 2012. Diversity, classification and function of the plant protein kinase superfamily. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **367**: 2619–2639.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**: 1754–1760.

Li HT, Yi TS, Gao LM, Ma PF, Zhang T, Yang JB, Gitzendanner MA, Fritsch PW, Cai J, Luo Y *et al.* 2019. Origin of angiosperms and the puzzle of the Jurassic gap. *Nature Plants* **5**: 461–470.

Li Z, Baniaga AE, Sessa EB, Scascitelli M, Graham SW, Rieseberg LH, Barker MS. 2015. Early genome duplications in conifers and other seed plants. *Science Advances* **1**: e1501084.

Li Z, Barker MS. 2020. Inferring putative ancient whole genome duplications in the 1000 Plants (1KP) initiative: access to gene family phylogenies and age distributions. *GigaScience* **9**: giaa004.

Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, De Smet R. 2016. Gene duplicability of core genes is highly consistent across all angiosperms. *Plant Cell* **28**: 326–344.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO *et al.* 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**: 289–293.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151–1155.

Madlung A. 2013. Polyploidy and its effect on evolutionary success: old questions revisited with new tools. *Heredity* **110**: 99–104.

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences, USA* **102**: 5454–5459.

Mayrose I, Zhan SH, Rothfels CJ, Magnuson-Ford K, Barker MS, Rieseberg LH, Otto SP. 2011. Recently formed polyploid plants diversify at lower rates. *Science* **333**: 1257–1257.

Monro AK, Bystriakova N, Fu L, Wen F, Wei Y. 2018. Discovery of a diverse cave flora in China. *PLoS ONE* **13**: e0190801.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**: 772–780.

Ohno S. 1970. *Evolution by gene duplication.* Berlin, Germany: Springer.

Oliver PM, Laver RJ, Martins FD, Pratt RC, Hunjan S, Moritz CC. 2017. A novel hotspot of vertebrate endemism and an evolutionary refugium in tropical Australia. *Diversity Distributions* **23**: 53–66.

Perez-Rodriguez P, Riano-Pachon DM, Correa LGG, Rensing SA, Kersten B, Mueller-Roeber B. 2010. PInTFDB: updated content and new features of the plant transcription factor database. *Nucleic Acids Research* **38**: D822–D827.

Proost S, Fostier J, Witte De, Dhoedt B, Demeester P, Van de Peer Y, Vandepoele K. 2012. i-ADHoRe 3.0-fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Research* **40**: e11.

Puglisi C, Yao TL, Milne R, Moller M, Middleton DJ. 2016. Generic recircumscription in the Loxocarpinae (Gesneriaceae), as inferred by phylogenetic and morphological data. *Taxon* **65**: 277–292.

Ren R, Wang HF, Guo CC, Zhang N, Zeng LP, Chen YM, Ma H, Qi J. 2018. Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. *Molecular Plant* **11**: 414–428.

Roalson EH, Roberts WR. 2016. Distinct processes drive diversification in different clades of Gesneriaceae. *Systematic Biology* **65**: 662–684.

Rushton PJ, Somssich IE, Ringler P, Shen QJ. 2010. WRKY transcription factors. *Trends in Plant Science* **15**: 247–258.

Sanderson MJ. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**: 662–684.

Schranz ME, Mohammadin S, Edger PP. 2012. Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. *Current Opinion in Plant Biology* **15**: 147–153.

Seoighe C, Gehring C. 2004. Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. *Trends in Genetics* **20**: 461–464.

Singh KB, Foley RC, Onate-Sanchez L. 2002. Transcription factors in plant defense and stress responses. *Current Opinion in Plant Biology* **5**: 430–436.

Smith SA, Brown JW, Yang Y, Bruenn R, Drummond CP, Brockington SF, Walker JF, Last N, Douglas NA, Moore MJ. 2018. Disparity, diversity, and duplications in the Caryophyllales. *New Phytologist* **217**: 836–854.

Sollars ESA, Harper AL, Kelly LJ, Sambles CM, Ramirez-Gonzalez RH, Swarbreck D, Kaithakottil G, Cooper ED, Uauy C, Havlickova L *et al.* 2017. Genome sequence and genetic diversity of European ash trees. *Nature* **541**: 212–216.

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng CF, Sankoff D, dePamphilis CW, Wall PK, Soltis PS. 2009. Polyploidy and angiosperm diversification. *American Journal of Botany* **96**: 336–348.

Soltis DE, Segovia-Salcedo MC, Jordon-Thaden I, Majure L, Miles NM, Mavrodiev EV, Mei WB, Cortez MB, Soltis PS, Gitzendanner MA. 2014. Are polyploids really evolutionary dead-ends (again)? A critical reappraisal of Mayrose et al. (2011). *New Phytologist* **202**: 1105–1117.

Soltis PS, Soltis DE. 2016. Ancient WGD events as drivers of key innovations in angiosperms. *Current Opinion in Plant Biology* **30**: 159–165.

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

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research* **34**: W609–W612.

Tank DC, Eastman JM, Pennell MW, Soltis PS, Soltis DE, Hinchliff CE, Brown JW, Sessa EB, Harmon LJ. 2015. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytologist* **207**: 454–467.

Unver T, Wu ZY, Sterck L, Turktas M, Lohaus R, Li Z, Yang M, He LJ, Deng TQ, Escalante FJ *et al.* 2017. Genome of wild olive and the evolution of oil biosynthesis. *Proceedings of the National Academy of Sciences, USA* **114**: E9413–E9422.

Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics* **10**: 725–732.

Vision TJ, Brown DG, Tanksley SD. 2000. The origins of genomic duplications in Arabidopsis. *Science* **290**: 2114–2117.

Wallace AR. 1858. On the tendency of varieties to depart indefinitely from the original type. *Journal of the Proceeding of the Linnean Society, Zoology* **3**: 53–62.

Wan T, Liu ZM, Li LF, Leitch AR, Leitch IJ, Lohaus R, Liu ZJ, Xin HP, Gong YB, Liu Y *et al.* 2018. A genome for gnetophytes and early evolution of seed plants. *Nature Plants* **4**: 82–89.

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

Wang J, Ai B, Kong HH, Kang M. 2017a. Speciation history of a species complex of *Primulina eburnea* (Gesneriaceae) from limestone karsts of southern China, a biodiversity hot spot. *Evolutionary Applications* **10**: 919–934.

Wang J, Feng C, Jiao TL, von Wettberg EB, Kang M. 2017b. Genomic signature of adaptive divergence despite strong nonadaptive forces on edaphic islands: a case study of *Primulina juliae*. *Genome Biology Evolution* **9**: 3495–3508.

Wang N, Yang Y, Moore MJ, Brockington SF, Walker JF, Brown JW, Liang B, Feng T, Edwards C, Mikenas J *et al.* 2019. Evolution of Portulacineae marked by gene tree conflict and gene family expansion associated with adaptation to harsh environments. *Molecular Biology Evolution* **36**: 112–126.

Wang S, Zhang JB, Jiao WQ, Li J, Xun XG, Sun Y, Guo XM, Huan P, Dong B, Zhang LL *et al.* 2017. Scallop genome provides insights into evolution of bilaterian karyotype and development. *Nature Ecology & Evolution* **1**: 0120.

Weber A, Clark JL, Moeller M. 2013. A new formal classification of Gesneriaceae. *Selbyana* **31**: 68–94.

Wei YG. 2010. *Gesneriaceae of South China.* Nanning, China: Guangxi Sciences and Technology Publishing House.

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences, USA* **106**: 13875–13879.

Wu S, Han B, Jiao Y. 2020. Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms. *Molecular Plant* **13**: 59–71.

Wu XL, Shiroto Y, Kishitani S, Ito Y, Toriyama K. 2009. Enhanced heat and drought tolerance in transgenic rice seedlings overexpressing OsWRKY11 under the control of HSP101 promoter. *Plant Cell Reports* **28**: 21–30.

Xu WB, Guo J, Pan B, Zhang Q, Liu Y. 2017. Diversity and distribution of Gesneriaceae in China. *Guihaia* **37**: 1219–1226.

Yang Y, Moore MJ, Brockington SF, Soltis DE, Wong GKS, Carpenter EJ, Zhang Y, Chen L, Yan ZX, Xie YL *et al.* 2015. Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. *Molecular Biology and Evolution* **32**: 2001–2014.

Yang Y, Smith SA. 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Molecular Biology and Evolution* **31**: 3081–3092.

Zhang Z, Xiao JF, Wu JY, Zhang HY, Liu GM, Wang XM, Dai L. 2012. ParaAT: A parallel tool for constructing multiple protein-coding DNA alignments. *Biochemical and Biophysical Research Communications* **419**: 779–781.

Zheng Y, Jiao C, Sun HH, Rosli HG, Pombo MA, Zhang PF, Banf M, Dai XB, Martin GB, Giovannoni JJ *et al.* 2016. iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Molecular Plant* **9**: 1667–1670.

Zwaenepoel A, Van de Peer Y. 2019. wgd-simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics* **35**: 2153–2155.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Fig. S1 Kmer frequency distribution of *P. huaijiensis.*

Fig. S2 Heat map of *P. huaijiensis* chromosome interaction.

Fig. S3 The distribution of GC content and sequencing depth in each 10-kbp window of *P. huaijiensis* genome.

Fig. S4 GO enrichment of tandem duplicates in *P. huaijiensis.*

Fig. S5 K_s distribution for orthologs from combinations between every two of the 18 asterids species.

Fig. S6 Syntenic relationship of self-comparison of the *P. huaijiensis* genome.

Fig. S7 GO enrichment of genes from all syntenic duplicates in *P. huaijiensis.*

Fig. S8 GO enrichment of duplicates from the *D* event in *P. huaijiensis*.

Fig. S9 GO enrichment of duplicates from the *L* event in *P. huaijiensis*.

Fig. S10 GO enrichment of *P. huaijiensis*-expanded genes.

Methods S1 Supplemental methods.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

Table S1 Statistics of the genome sequencing data of *P. huaijiensis*.

Table S2 Statistics of the *P. huaijiensis* genome assembly.

Table S3 Evaluation of the *P. huaijiensis* genome assembly using remapping of reads from short-insert libraries.

Table S4 Evaluation of the *P. huaijiensis* genome assembly using CEGMA.

Table S5 Evaluation of the *P. huaijiensis* genome assembly using BUSCO.

Table S6 Evaluation of the *P. huaijiensis* genome assembly using EST data.

Table S7 Statistics of the *P. huaijiensis* RNA-Seq data from different tissues and development stages.

Table S8 Summary of transposable elements in *P. huaijiensis*.

Table S9 Summary of gene models in *P. huaijiensis*.

Table S10 Summary of protein-coding gene annotation of *P. huaijiensis*.

Table S11 Summary of RNA-Seq data from eight subtribe Didymocarpinae species.

Table S12 Transcription factors in 34 sequenced eudicot genomes.