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

Chao Feng¹, Jing Wang¹, Lingqing Wu², Hanghui Kong¹, Lihua Yang¹, Chen Feng¹, Kai Wang², Mark Rausher³ and Ming Kang¹,⁴

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

doi: 10.1111/nph.16588

**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.

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

**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 karst 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; Solis 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; Solis et al., 2009; Jiao et al., 2010).
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, Edget et al. (2015) found increased diversification rates following ancient WGDs in Brassiccales. 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 Peet 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/scbgfe ngchao/; 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/geo), we assembled the P. huaijiensis genome according to a hybrid-specific SOAPdenovoovo 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 pseudomolecules 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...
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. huajiensis and 16 other model sequenced plants, including Tectona grandis, Handroanthus impetiginosus, Sesamum indicum, Antirrhinum majus, Dorcopsis hygrometricum (original name: Bœa hygrometrica; Puiglesi et al., 2016), Olea europaea, Fraxinus excelsior, Solanum tuberosum, Solanum lycopersicum, Aratibopsis italiana, Theobroma cacao, Populus trichocarpa, Vitis vinifera, Orniza 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.6.664b) (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.

We then 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. huajiensis, D. hygrometricum, O. europaea, F. excelsior and S. lycopersicum) and eight Didymocarpaceae transcripts. In Dataset 2, we added two species (S. lycopersicum and V. vinifera) to Dataset 1.

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. huajiensis, D. hygrometricum, O. europaea, F. excelsior and S. lycopersicum) and eight Didymocarpaceae transcripts. In Dataset 2, we added two species (S. lycopersicum and V. vinifera) to Dataset 1.

We identified homologs and orthologs from both Datasets 1 and 2 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 obtained transcriptomes from eight subtribe Didymocarpaceae species, including Henkelia anchoreta, Cyrtandra dispar, Hemiboea subcapitata, Petrocodon fangianus, Primulina rubella, Primulina swingei, Primulina fimbriatepa 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.

Transcriptome assembly

We obtained transcriptomes from eight subtribe Didymocarpaceae species, including Henkelia anchoreta, Cyrtandra dispar, Hemiboea subcapitata, Petrocodon fangianus, Primulina rubella, Primulina swingei, Primulina fimbriatepa 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. huajiensis, D. hygrometricum, O. europaea, F. excelsior and S. lycopersicum) and eight Didymocarpaceae transcripts. In Dataset 2, we added two species (S. lycopersicum and V. vinifera) to Dataset 1.

We used Orthofinder v.2.3.3 (Emms & Kelly, 2015) with the parameter (-S diamond -og) to classify the orthogroups of proteins from P. huajiensis and 16 other model sequenced plants, including Tectona grandis, Handroanthus impetiginosus, Sesamum indicum, Antirrhinum majus, Dorcopsis hygrometricum (original name: Bœa hygrometrica; Puiglesi et al., 2016), Olea europaea, Fraxinus excelsior, Solanum tuberosum, Solanum lycopersicum, Aratibopsis italiana, Theobroma cacao, Populus trichocarpa, Vitis vinifera, Orniza 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.6.664b) (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 script (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.
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/GG PLOT2, 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 1-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 PALEOPOLOIDY 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 the tree species (((((*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/mcsan2/) 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 (Figsaw doi: 10.6084/m9.figshare.10185056). We then drew the 4DTv distribution and labeled the peak using R/GG PLOT2, 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/tangha ibao/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.org/packages/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 PlnTDB 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: $z = \frac{(x - \mu)}{\sigma}$, where $x$, $\mu$ and $\sigma$ 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 220 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 Mbp (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).

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 transerase 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 transcripts (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).

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

<table>
<thead>
<tr>
<th>Assembly feature</th>
<th>Number (percentage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome-sequencing depth (×)</td>
<td>530</td>
</tr>
<tr>
<td>Estimated genome size (Mbp)</td>
<td>511</td>
</tr>
<tr>
<td>Total length of scaffolds (Mbp)</td>
<td>478</td>
</tr>
<tr>
<td>N50 of scaffolds (bp)</td>
<td>23 479 473</td>
</tr>
<tr>
<td>Total Length of contigs (Mbp)</td>
<td>466</td>
</tr>
<tr>
<td>N50 of contigs (bp)</td>
<td>28 983</td>
</tr>
<tr>
<td>Mapping rate by reads from short-insert libraries</td>
<td>98.5%</td>
</tr>
<tr>
<td>Core Eukaryotic Genes Mapping Approach (CEGMA evaluation)</td>
<td>97.2%</td>
</tr>
<tr>
<td>Benchmarking Universal Single-Copy Orthologs (BUSCO evaluation)</td>
<td>96.5%</td>
</tr>
<tr>
<td>EST evaluation</td>
<td>94.5%</td>
</tr>
<tr>
<td>RNA-Seq evaluation</td>
<td>89.0–93.2%</td>
</tr>
<tr>
<td>Genome annotation</td>
<td></td>
</tr>
<tr>
<td>Percentage of transposable elements (TE)</td>
<td>54.1%</td>
</tr>
<tr>
<td>Percentage of long terminal repeat-retrotransposons (LTR)</td>
<td>48.4%</td>
</tr>
<tr>
<td>No. of predicted protein-coding genes</td>
<td>31 328</td>
</tr>
<tr>
<td>No. of genes annotated to public database</td>
<td>30 583 (97.6%)</td>
</tr>
<tr>
<td>No. of genes annotated to GO database</td>
<td>18 781 (59.9%)</td>
</tr>
<tr>
<td>No. of genes duplicated by tandem duplications</td>
<td>19 48 (6.2%)</td>
</tr>
<tr>
<td>No. of genes duplicated by syntenic duplications</td>
<td>15 197 (48.5%)</td>
</tr>
<tr>
<td>No. of genes duplicated by the D event</td>
<td>10 132 (32.3%)</td>
</tr>
</tbody>
</table>

GO, gene ontology.
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 Ks 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 orthologs. (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 orthologs. 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 divergence time was estimated on a basis of three calibration points (blue circles). CI, confidence interval. The numbers in red and green indicate the age (Ma) from the K–Pg boundary to the present. (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 + green + blue bars, genes belonging to 6496 asterid-shared orthogroups in each of ten asterids; gray + 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 10791 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.

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 10132 (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 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 Primulina 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).
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 $P < 1E-30$; (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).

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

<table>
<thead>
<tr>
<th>GO ID</th>
<th>Term annotated</th>
<th>The D event</th>
<th>The L event</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of background</td>
<td>No. of genes</td>
<td>$P$-value$^a$</td>
</tr>
<tr>
<td>GO:0043565</td>
<td>Sequence-specific DNA binding</td>
<td>542</td>
<td>244</td>
</tr>
<tr>
<td>GO:0003682</td>
<td>Chromatin binding</td>
<td>418</td>
<td>169</td>
</tr>
<tr>
<td>GO:0003700</td>
<td>DNA binding transcription factor activity</td>
<td>823</td>
<td>301</td>
</tr>
<tr>
<td>GO:0004713</td>
<td>Protein tyrosine kinase activity</td>
<td>1186</td>
<td>452</td>
</tr>
<tr>
<td>GO:0008270</td>
<td>Zinc ion binding</td>
<td>1285</td>
<td>525</td>
</tr>
<tr>
<td>GO:0015291</td>
<td>Secondary active transmembrane transporter activity</td>
<td>121</td>
<td>55</td>
</tr>
<tr>
<td>GO:0004842</td>
<td>Ubiquitin-protein transferase activity</td>
<td>117</td>
<td>54</td>
</tr>
<tr>
<td>GO:0005515</td>
<td>Protein binding</td>
<td>3991</td>
<td>1549</td>
</tr>
<tr>
<td>GO:0016616</td>
<td>Oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor</td>
<td>162</td>
<td>72</td>
</tr>
<tr>
<td>GO:0003735</td>
<td>Structural constituent of ribosome</td>
<td>443</td>
<td>155</td>
</tr>
<tr>
<td>GO:0098772</td>
<td>Molecular function regulator</td>
<td>228</td>
<td>79</td>
</tr>
<tr>
<td>GO:0016879</td>
<td>Ligase activity, forming carbon-nitrogen bonds</td>
<td>141</td>
<td>58</td>
</tr>
</tbody>
</table>

$^a$The $P$ values $< 0.01$ are in bold.

**Prefential 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
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-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.

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

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

aThe percentage of transcription factors (TF) in categories of expanded genes with specific GO terms; bThe 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-6) from specific duplications were shown in details around the corresponding square.
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 Didymocarpiniae 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’ (Schrantz et al., 2012), where increase in diversification rates tend to follow WGD events after a lag time of millions of years (Schrantz 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), Gesnerioidae (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, Didymocarpiniae 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 Cyrtandra (the largest genus with approx. 800 species), Primulina (approx. 180 species), Aeschynanthus (approx. 160 species) and Oreocharis (c. 140 species), belong to Didymocarpiniae, and experienced a shared D event. Roalson & Roberts (2016) identified elevated diversification rates in several lineages in Didymocarpiniae, 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 Didymocarpiniae.

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 non-random 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 huaianensis experienced two rounds of WGD after the $\gamma$ 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. huaianensis (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 adaptation 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 hatleri*ana, *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 (Ca2+) 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 Ca2+ and magnesium ion (Mg2+)) 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-based, tree-based and syntenic approaches showed that *P. huaijiensis* experienced two rounds of WGD since the γ triplation 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 (bZIPs) 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


### 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* 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.