iScience, Volume *23*

Supplemental Information

Genome Sequencing of the Endangered *Kingdonia uniflora* (Circaeasteraceae, Ranunculales) Reveals Potential Mechanisms of Evolutionary Specialization

Yanxia Sun, Tao Deng, Aidi Zhang, Michael J. Moore, Jacob B. Landis, Nan Lin, Huajie Zhang, Xu Zhang, Jinling Huang, Xiujun Zhang, Hang Sun, and Hengchang Wang

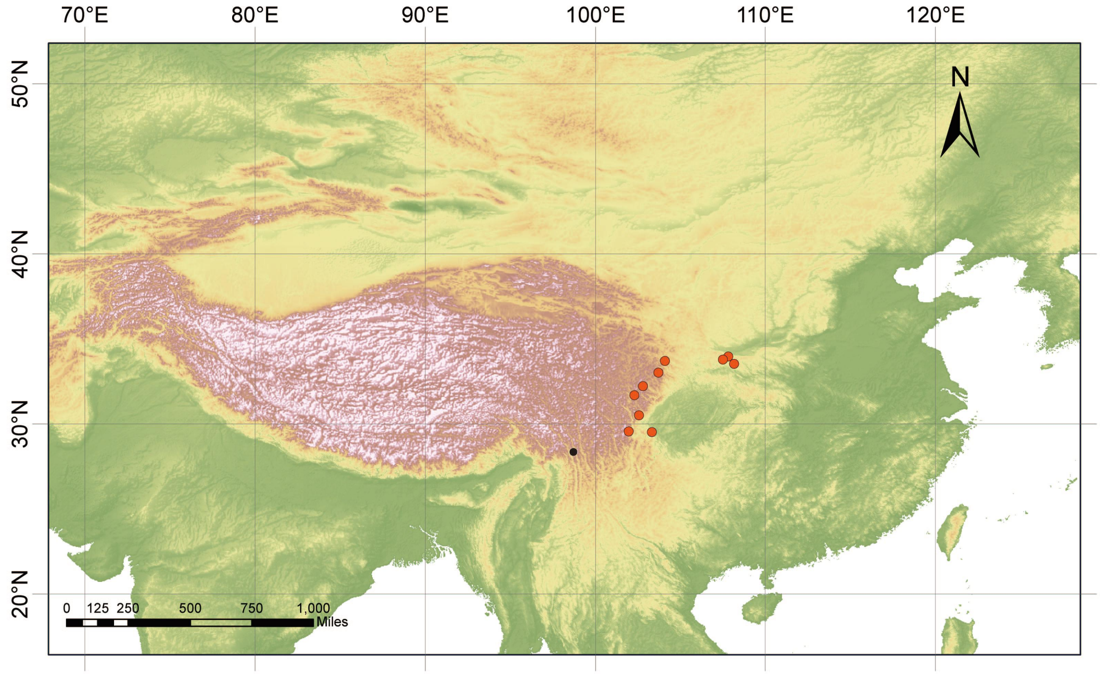

Figure S1. Distribution range of *K. uniflora.* The black dot shows the individuals with previous occurrence record but no longer have extant populations; the red dots represent the current distribution range of *K. uniflora*. Related to Table 1

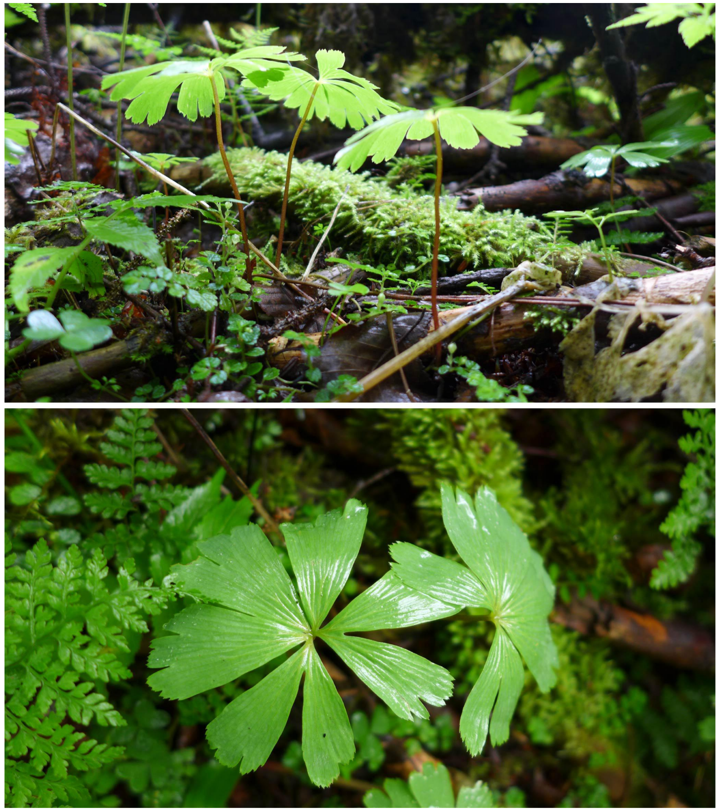

Figure S2. Morphological features of *K. uniflora.* Related to Table 1.

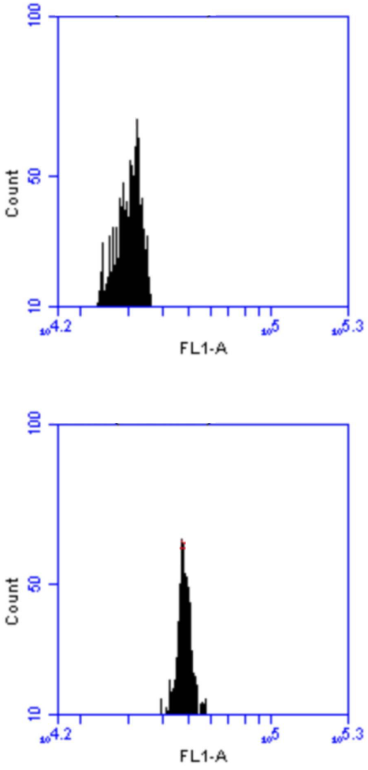

Figure S3. Estimation of *K. uniflora* genome size based on flow cytometer analysis. The above panel showing the 2C DNA of *Actinidia chinensis* (Hopping, 1994) at 32377.78, and the panel below indicating 2C DNA of *K. uniflora* at 47614.17. Related to Table 1.

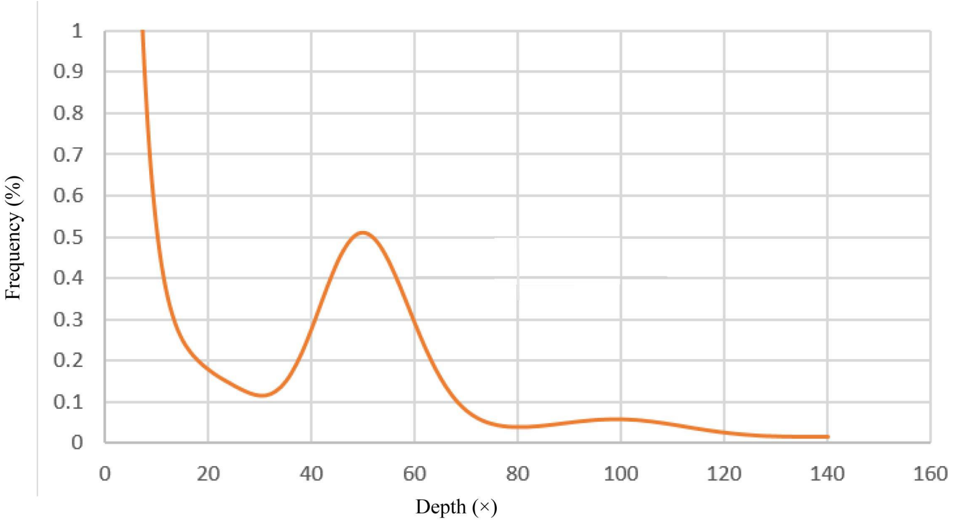

Figure S4. 17 *k*-mer frequency distribution of sequencing reads. Related to Table 1.

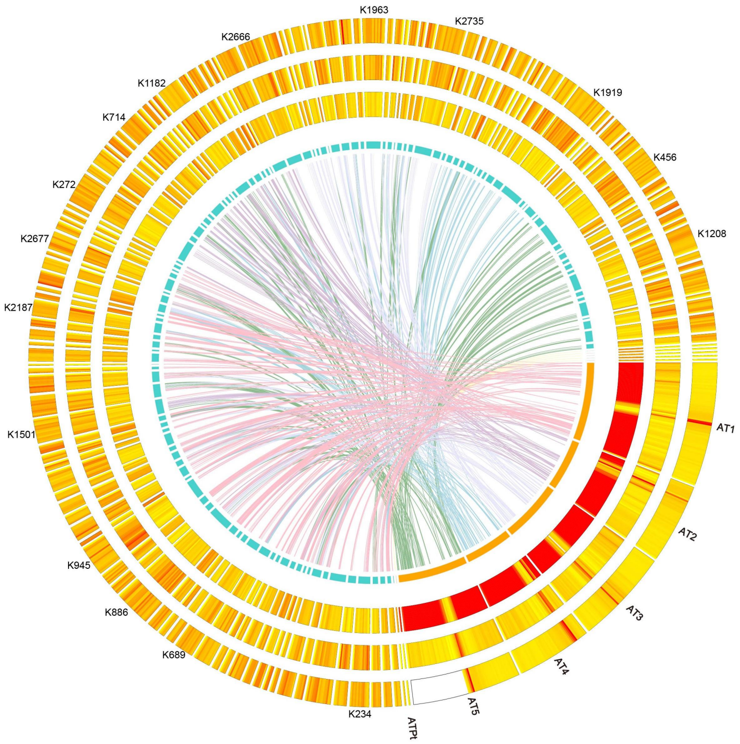

Figure S5. Comparative analyses of genomic features between *Kingdonia uniflora* and *Arabidopsis thaliana*. Tracks from inside to outside are collinearity between both genomes, number of chromosomes/scaffolds, gene density, GC content and TE density. Related to Figure 1.

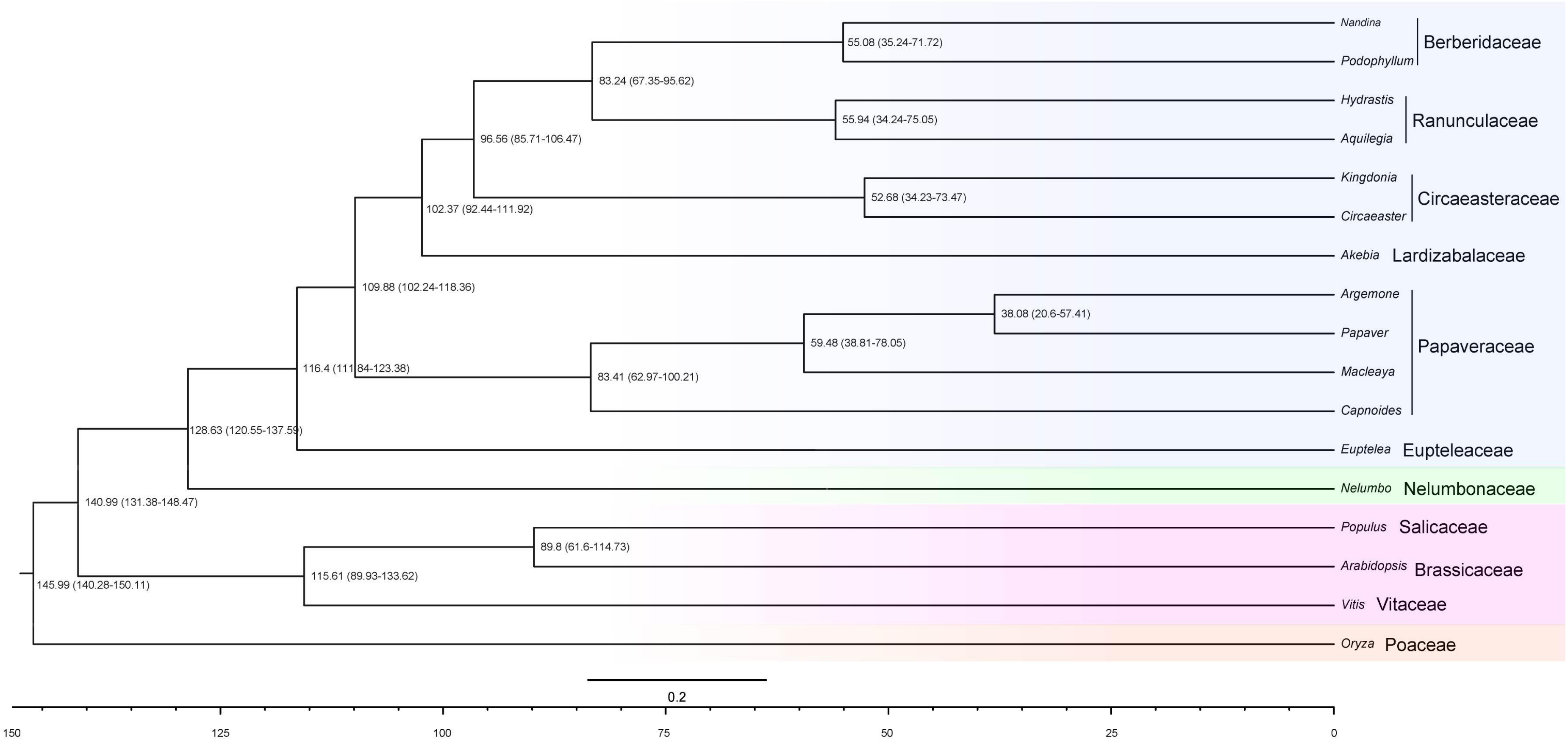

Figure S6. Dated phylogeny for 17 plant species with *Oryza* as an outgroup. A time scale is shown at the bottom. Related to Figure 3.

Table S1. Statistics of characteristics of *K. uniflora* genome (*K*-mer=17). Related to Table 1.

Chracteristics	
K-mer	17
Peak_Depth	80
N *K*-mer	58,788,492,58
Genome size	1170 Mb
Revised Genome size	1005 Mb
Heterozygous rate	nan
Repeat rate	0.6242

Table S2. Sequencing and quality filtering statistics. Related to Table 1.

	Total data (G)	Sequence coverage (X)
Illumina sequencing	236	201.7
Pacbio Sequencing	106.5	90.6
Total	342.5	292.7

Table S3. Information of function annotation in *K. uniflora* genes. Related to Table 1.

Database	Annotated Number	Annotated Percent (%)
NR	35818	82.7%
Swiss-Prot	27859	64.3%
KEGG	32104	74.1%
InterPro	25951	59.9%
Total	35953	83.03%

Table S4. Statistics of noncoding RNA in *K. uniflora* genome. Related to Table 1.

Type		Copy	Average length (bp)	Total length (bp)
tRNA		1124	76	85487
rRNA		715		
	18S	81	1802	146008
	28S	76	6732	511650
	5.8S	0	_	_
	5S	558	115	63990
snRNA	CD-box	1447	104	149886
	HACA-box	97	124	11991
	splicing	207	154	31895
miRNA		125	126	15757

Table S5. Scaffolds from the *K. uniflora* assembly were aligned to conserved genes using BUSCO method. Related to Table 1.

Species	Genome Size	BUSCO annotation assessment results
K. uniflora	1004.7 Mb	C:90.6% [D:4.1%], F:3.3%, M:6.1%, n:1375

C: Complete Single-Copy BUSCOs

D: Complete Duplicated BUSCOs

F: Fragmented BUSCOs

M: Missing BUSCOs

n: Total BUSCO groups searched

Table S6. Comparison of genome assembly within Ranunculales. Related to Table 1.

Species	Family	Genome size (Mb)	No. of scaffolds	No. of contigs	Length of N50, bp	References
Aquilegia coerulea	Ranunculaceae	301.98	970	7,189	121,821	Filiault et al., 2018
Berberis thunbergii	Berberidaceae	2240.74	11,815	11,815	397,058	NCBI available
Kingdonia uniflora	Circaeasteraceae	1004.7	2,932	2932	2,099,369	
Eschscholzia californica	Papaveraceae	489.065	53,253	85,931	20,647	Hori et al., 2018
Macleaya cordata	Papaveraceae	377.834	4,547	25,550	36,130	Liu et al., 2017
Papaver somniferum	Papaveraceae	2715.53	34,381	65,344	1,773,300	Guo et al., 2018

Table S7. Statistics of repeat sequences and transposable elements in *K. uniflora* genome. Related to Table 1.

Type	Repeat size	Percent of genome (%)
DNA	50812195	5.06
LINE	30109270	3.00
SINE	173746	0.02
LTR	408124656	40.62
Simple repeat	5211550	0.52
Unknown	179205100	17.84
Total	671481635	66.83

Table S9. Length comparasion of *ndh* genes between *K. uniflora* and *C. agrestis.* Related to Figure 4.

	C. agrestis (bp)	*K. uniflora* (bp)
ndhA	1089	553
ndhB	1530	723
ndhC	360	0
ndhD	1500	15
ndhE	303	303
ndhF	2205	0
ndhG	528	0
ndhH	1179	615
ndhI	540	0
ndhJ	474	468
ndhK	666	234

Table S10. Information of species used for phylogenetic analyses. Colored characters showing the expanded taxa in Figure S5 compared with that in Figure 3. Related to Figure 3.

Species	Family	Source
Oryza sativa L.	Poaceae	NCBI
Vitis vinifera L.	Vitaceae	NCBI
Populus trichocarpa Torr. and Gray	Salicaceae	NCBI
Arabidopsis thaliana L.	Brassicaceae	NCBI
Nelumbo nucifera Gaertner	Nelumbonaceae	NCBI
Euptelea pleiosperma J. D. Hooker and Thomson	Eupteleaceae	1 kp
Argemone mexicana L.	Papaveraceae	1 kp
Papaver bracteatum Lindl.	Papaveraceae	1 kp
Capnoides sempervirens (L.) Borkh.	Papaveraceae	1 kp
Macleaya cordata (Willd.) R. Br.	Papaveraceae	NCBI
Akebia trifoliata (Thunberg) Koidzumi	Lardizabalaceae	1 kp
Circaeaster agrestis Maxim.	Circaeasteraceae	Current study
Kingdonia uniflora Balf. f. and W.W. Sm.	Circaeasteraceae	Current study
Hydrastis canadensis L.	Ranunculaceae	1 kp

Aquilegia coerulea E. James	Ranunculaceae	NCBI
Podophyllum peltatum L.	Berberidaceae	1 kp
Nandina domestica Thunberg	Berberidaceae	1 kp

Transparent Methods

1 Genome sequencing and assembly

1.1 Plant materials and sequencing

Fresh *K. uniflora* leaves were collected from individuals growing from the same rhizome in the Taibai Mountains (altitude 2,844 m, N 34.038°, E107.715°), Shaanxi, China. Total genomic DNA (≥10 ug, ≥50 ng/ul) was isolated from fresh leaves using the conventional cetyltriethylammonium bromide (CTAB) method (Doyle and Doyle, 1987). To help to estimate the genome size and polish genome assembly, we conducted Illumina sequencing; two paired-end sequencing libraries with insert sizes of 270 bp and 500 bp, respectively, were constructed and sequenced on the Illumina HiSeq X ten platform (Illumina Inc., CA, USA) at Beijing Genomics Institute (BGI) in Wuhan, Hubei, China. For PacBio single-molecule real-time sequencing, sequencing libraries with 20-kb DNA inserts were constructed and sequenced on the PacBio Sequel platform (Pacific Biosciences, CA, USA) at BGI. We also collected fresh leaves of *C. agrestis* in Taibai Mountains (altitude 2,837m, N34.038°, E107.68) for RNA extraction. Total RNA was extracted from young leaves (~100 mg) of both *K. uniflora* and *C. agrestis* using TRIzol Reagent RNA Purification (DSB, Guangdong, China). A cDNA library with insert sizes of 350-400 bp was prepared using NEBNext Ultra RNA Library Prep Kit for Illumina (NEB, MA, USA) and paired-end sequenced on the HiSeq X ten platform (Illumina Inc., CA, USA) at

BGI. The transcriptome data of *K. uniflora* was used for the prediction of protein-coding genes; and the transcriptome data of *C. agrestis* was used to detect single copy genes.

1.2 *De novo* assembly

The PacBio long reads were first error corrected and then *de novo* assembled using Canu v1.8 (Koren et al., 2017) with default parameters (rawErrorRate=0.300, correctedErrorRate=0.045, minReadLength=1000, minOverlapLength=500, canuIterationMax=2) except for setting the genome size to 1.2 G to obtain contigs. Then iterative polishing was conducted on the Canu derived contigs using Pilon v1.2.3 (Walker et al., 2014) in which adapter-trimmed paired-end Illumina reads from DNA sequencing were aligned with the raw assembly with default parameters to fix bases and correct local misassembles. RNA-seq reads were assembled into transcripts using Trinity v2.6.6 (Grabherr et al., 2011) with the paired-end option and remaining default parameters.

2 Genome annotation

2.1 Annotation of repetitive sequences

We identified *de novo* repetitive sequences in the *K. uniflora* genome using RepeatModeler (http://www.repeatmasker.org/RepeatModeler/) based on a self-blast search. We further used RepeatMasker (http: //www.repeatmasker.org/) to search for known repetitive sequences using a cross-match program with a Repbase-derived

RepeatMasker library and the *de novo* repetitive sequences constructed by RepeatModeler. Intact LTR (long terminal repeat) retrotransposons were identified by searching the genome of *K. uniflora* with LTRharvest (Ellinghaus et al., 2008) (-motif tgca -motifmis 1) and LTR_Finder (Xu and Wang, 2007) (-D 20000 -d 1000 -L 5000 -I 100). We combined results from both analyses and filtered false positives using LTR_retriever (Qu and Jiang, 2018), which also calculated the insertion date (*t*) for each LTR retrotransposons (*t= K/2r*, K: genetic distance) using a substitution rate (*r*) of 1.4*10−9 substitutions per site per year calculated by MCMCtree in PAML (Yang, 2007) .

2.2 Structural and functional annotation of genes

Putative protein-coding gene structures in the *K. uniflora* genome were homology predicted using the Maker package v2.31.10 (Holt and Yandell, 2011) with protein references from the published Ranunculales genomes and the *de novo* assembled transcripts of *K. uniflora* transcriptome data generated in this study, and *de novo* predicted using Augustus v3.3.2 (Stanke et al., 2006). The rRNAs were predicted using RNAmmer v1.2 (Lagesen et al., 2007), tRNAs were predicted using tRNAscan-SE v1.4 (Lowe and Eddy, 1997), and other noncoding RNA sequences were identified using Rfam v12.0 by inner calling using Infernal v1.1.2 (Nawrocki and Eddy, 2013).

Functional annotation of the protein-coding genes was carried out by performing BLASTP analyses (e-value cut-off 1e-05) against the NCBI nonredundant protein sequence database and SwissProt. Searches for gene motifs and domains were performed using InterProScan v5.16.55

(Jones et al., 2014). Completeness of the genome was assessed by performing gene annotation using the BUSCO (v3.0.2) method (Simão et al, 2015) by searching the Embryophyta library.

3 Investigation of whole-genome duplication

We identified paralogs (within *K. uniflora* and *C. agrestis*, respectively) and orthologs (between *K. uniflora* and *C. agrestis*) using BLASTP (E value = 1E-07). For each gene pair, the number of synonymous substitutions per synonymous site (*Ks*) based on the NG method was calculated using TBtools (Chen et al., 2018); *Ks* values of all gene pairs were plotted to identify putative whole-genome duplication events. In addition, MCScanx (Wang et al., 2012) was used to identify syntenic blocks within the *K. uniflora* genome. Dot-plot analysis of syntenic blocks with at least five gene pairs was conducted using the dot plotter program within the MCScanX package to further detect whole-genome duplication events.

4 Gene family and phylogenomic analysis

Orthogroups were constructed with 11 other sequenced plant species (Table S10). To get reliable tree topology, all basal eudicot species (*Nandina domestica, Aquilegia coerulea, Circaeaster agrestis, Akebia trifoliata, Macleaya cordata, Euptelea pleiosperma,* and *Nelumbo nucifera*) with available genome sequences were included for the phylogenetic analyses. In addition, the frequently-used outgroups (*Arabidopsis thaliana, Populus trichocarpa, Vitis vinifera* and *Oryza sativa*) in previous studies (e.g., Chen et al., 2018; Song et al., 2018; Yang et al., 2019) were also included in our analyses. CD-HIT (Huang et al., 2010) was employed to remove redundancy caused by alternative splicing variations (-c 0.8 -aS 0.8). To exclude putative fragmented genes, genes encoding protein sequences shorter than 50 aa (amino acids) were filtered out. All filtered protein sequences of the 12 species were compared with each other using BLASTP (E value = 1E-5) and clustered into orthologous groups by OrthoFinder (Emms and Kelly, 2015). Protein sequences of single-copy gene families identified by OrthoFinder were used for phylogenetic tree construction. MAFFT version 7.0 (Katoh and Standley, 2013) was used to generate multiple sequence alignment for protein sequences in each single-copy family. Poorly aligned regions were further trimmed using the Gblocks (Castresana, 2000; Talavera and Castresana, 2007). The alignments of each gene family were concatenated to a super alignment matrix, which was then used for phylogenetic tree reconstruction through the PROTCATJTT model in RAxML version 8.1.2 (Stamatakis, 2014). To assess species tree clade support, a coalescent-based analysis was also conducted using RAxML bootstrap gene trees as input for ASTRAL v. 4.7.6 (Mirarab et al., 2015). To test the accuracy of above phylogenetic analyses, a second data set consisting of 17 taxa was also used following the same steps which consisted of increased taxonomic sampling with the tradeoff of fewer loci.

To investigate the evolutionary history of *K. uniflora*, divergence time between 12 species was estimated using MCMCtree in PAML (Yang, 2007) with the options “independent rates” and “HKY85” model. A Markov chain Monte Carlo analysis was run for 100,000,000 generations, using a burn-in of 1,000 iterations. Two constraints were used for time calibrations: (1) 140–150 Mya for the monocot-dicot split (Gaut et al., 1996; Yang et al., 2018); 112-124 Mya for the Ranunculales crown group (Magallón et al., 2015; Sun et al., 2018).

5 Gene family overrepresentation and underrepresentation

Overrepresentation and underrepresentation of the OrthoFinder-derived orthologous gene families were determined using CAFÉ v. 4.1 (De Bie et al., 2006). The program uses a birth and death process to model gene gain and loss across a user-specified phylogenetic tree. The distribution of family sizes generated under this model can provide a basis for assessing the significance of the observed family size differences among taxa. For each significantly overrepresented and underrepresented gene family in *K. uniflora*, functional information was inferred via KOBAS (http://kobas.cbi.pku.edu.cn/anno_iden.php) using KEGG Pathway database.

6 Plastid *ndh* gene searching

To examine whether the plastid *ndh* genes have been completely lost from *K. uniflora* or transferred to the nuclear genome, intact **s**equences of all (11) plastid *ndh* genes (*ndhA*, *ndhB*, *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ* and *ndhK*) were extracted from the plastome of *C. agrestis* (Sun et al., 2017). Then BLASTN analyses (E value = 1E-5) between the 11 gene sequences and assembled *K. uniflora* genome sequences was conducted.

Supplemental references

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. *17,* 540-552.

Chen, C., Chen, H., He, Y.H., and Xia, R. TBtools, a Toolkit for Biologists integrating various biological data handling tools with a user-friendly interface. DOI: https://doi.org/10.1101/289660 (2018)

Chen, J., Hao, Z., Guang, X., Zhao, C., Wang, P., Xue, L., Zhu, Q., Yang, L., Sheng, Y., Zhou, Y., et al. (2019). *Liriodendron* genome sheds light on angiosperm phylogeny and species-pair differentiation. Nat. Plants *5,* 18-25.

De Bie, Cristianini, N., Demuth, J.P., and Hahn, M.W. (2006). CAFÉ: a computational tool for the study of gene family evolution. Bioinformatics *22,* 1269-1271.

Doyle, J.J., and Doyle, J.L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. 19, 11-15.

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics *9,* 18.

Emms, D.M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. *16,* 157.

Gaut, B.S., Morton, B.R., McCaig, B.C., and Clegg, M.T. (1996). Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene *rbcL*. Proc. Natl. Acad. Sci. *93,* 10274-10279.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al.

(2011). Full-length transcriptome without a genome from RNA-Seq data. Nat. Biotechnol. *29,* 644-652.

Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics *12,* 491.

Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics *26,* 680.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30,* 1236-1240.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30,* 772-780.

Koren, S. Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. *27,* 722-736.

Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H.-H., Rognes, T., and Ussery, D.W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. *35,* 3100-3108.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. *25,* 955-964.

Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L.L., and Hernández-Hernández, T. (2015). A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. New Phytol. *207,* 437-453.

Mirarab, S., and Warnow, T. (2015). ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics *31,* 44–52.

Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29:, 2933-2935.

Qu, S.J., and Jiang, N. (2018). LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. Plant Physiol. *176,* 1410-1422.

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

Song, C., Liu, Y., Song, A., Dong, G., Zhao, H., Sun, W., Ramakrishnan, S., Wang, Y., Wang, S., Li, T., et al. (2018). The Chrysanthemum nankingense Genome Provides Insights into the Evolution and Diversification of Chrysanthemum Flowers and Medicinal Traits. Mol Plant *11,* 1482-1491.

Stamatakis, A., Ludwig, T., and Meier, H. (2004). RAxML-III: A fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics *21,* 456-463.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. *34,* W435-W439.

Sun, Y., Moore, M. J., Landis, J. B., Lin, N., Chen, L., Deng, T., Zhang, J.W., Meng, A.P., Zhang, S.J., Tojibaev, O.S., et al. (2018). Plastome phylogenomics of the early-diverging eudicot family Berberidaceae. Mol. Phylogenet. Evol. *128,* 203-211.

Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic Biol. *56,* 564-577.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One *9,* e112963.

Wang, E., Kirby, E., Furlong, K. P., van Soest, M., Xu, G., Shi, X., Kamp, P.J.J., and Hodges, K. V. (2012). Two-phase growth of high topography in eastern Tibet during the Cenozoic. Nat. Geosci. *5,* 640-645.

Xu, Z., and Wang, H. (2007). LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. *35,* W265-W268.

Yang, X., Yue, Y., Li, H., Ding, W., Chen, G., Shi, T., Chen, J., Park, M.S., Chen, F., and Wang, L. (2018). The chromosome-level quality genome provides insights into the evolution of the biosynthesis genes for aroma compounds of *Osmanthus fragrans*. Horticulture Research *5,* 72.

Yang, Y., Ma, T., Wang, Z., Lu, Z., Li, Y., Fu, C., Chen, X., Zhao, M., Olson, M.S., and Liu, J. (2018). Genomic effects of population collapse in a critically endangered ironwood tree *Ostrya rehderiana*. Nature Communications *9,* 5449.

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24,* 1586-1591.