


## ORIGINAL RESEARCH

# Draft genome of the famous ornamental plant *Paeonia suffruticosa*

Shuzuo Lv<sup>1,3</sup> | Shu Cheng<sup>2,5</sup>  | Zhanying Wang<sup>1,3</sup> | Shiming Li<sup>2,5</sup> | Xin Jin<sup>2,5</sup> | Lei Lan<sup>2,5</sup> | Bing Yang<sup>2,5</sup> | Kang Yu<sup>2,5</sup> | Xuemei Ni<sup>2,5,6,7</sup> | Ning Li<sup>2,3,5,6,7</sup> | Xiaogai Hou<sup>4</sup> | Gang Huang<sup>2,5,6,7</sup> | Jie Wang<sup>2,5,6,7</sup> | Yang Dong<sup>2,3,5,6,7</sup> | Erqiang Wang<sup>1,3</sup> | Jiangtao Huang<sup>1,3</sup> | Gengyun Zhang<sup>2,3,6,7</sup> | Canjun Zhang<sup>1,3</sup>

<sup>1</sup>Luoyang Academy of Agricultural and Forestry Sciences, Luoyang, Henan, China

<sup>2</sup>BGI-Shenzhen, Shenzhen, China

<sup>3</sup>BGI-Luoyang Agricultural innovation center, Luoyang, Henan, China

<sup>4</sup>College of Agriculture, Henan University of Science and Technology, Luoyang, Henan, China

<sup>5</sup>BGI Institute of Applied Agriculture, BGI-Shenzhen, Shenzhen, China

<sup>6</sup>State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen, China

<sup>7</sup>Key Laboratory of Genomics, Ministry of Agriculture, BGI-Shenzhen, Shenzhen, China

**Correspondence**

Canjun Zhang, Gengyun Zhang and Jiangtao Huang, BGI-Luoyang Agricultural innovation center, Luoyang, 471000, Henan, China. Emails: lynkyzcj@126.com (CZ); zhanggengyun@genomics.cn (GZ); lynkyhjt@126.com (JH)

**Funding information**

Luoyang Municipal Government, China; Shenzhen Municipal Government, China, Grant/Award Number: JCYJ20150831201123287 and JCYJ20160331150844452; Shenzhen, Grant/Award Number: JCYJ20160331150844452 and JCYJ20150831201123287

**Abstract**

Tree peony (*Paeonia* Sect. *Moutan*) is a famous ornamental plant, with huge historical, cultural, and economic significance worldwide. In this study, we reported the ~13.79 Gb draft genome of a wide-grown *Paeonia suffruticosa* cultivar "Luo shen xiao chun," representing the largest sequenced genome in dicots to date. Phylogenetic analyses based on genome sequences demonstrated that *P. suffruticosa* was placed as sister to Vitales, and they together formed a clade that was sister to Rosids, weakly supporting a relationship of ((Saxifragales and Vitales) and Rosids). The identification and expression analysis of MADS-box genes based on the genome assembly and de novo transcriptome assembly of *P. suffruticosa* revealed that the function of C class genes was restricted in flower development, which might be responsible for the stamen petalody in tree peony cultivars. Overall, the first sequenced genome in the family Paeoniaceae provides an important resource for the origin, domestication, and evolutionary study as well as cultivar breeding in tree peony.

**KEYWORDS**

Comparative genomics, draft genome, MADS-box, *Paeonia suffruticosa*, Tree peony

Lv, Cheng and Wang contributed equally to this work.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *Ecology and Evolution* published by John Wiley & Sons Ltd.

## 1 | INTRODUCTION

The genus *Paeonia* is well known for its high ornamental and medical values. It is the only genus in the family Paeoniaceae and consists of 33 species which are assigned to three sections: *Moutan*, *Paeonia*, and *Onaepia* (Christenhusz & Byng, 2016; Ji, Wang, Teixeira da Silva, & Yu, 2012). The tree peonies (*Paeonia* Sect. *Moutan*), native to China, have a long history of cultivation for over 1,600 years (Li, Zhang, & Zhao, 2011). They are perennial deciduous shrubs and were crowned the “king of flowers” for their large and varying forms of flowers, rich and bright colors, symbolizing happiness, wealth, and prosperity in Chinese culture. The seeds of the tree peonies contain rich unsaturated fatty acids such as  $\alpha$ -linolenic acid, oleic acid, and linoleic acid, and are considered to be a novel resource of high-value edible oil (Li, Wang, et al., 2015a; Li, Yuan, et al., 2015b). *Paeonia suffruticosa* belongs to Section *Moutan*, comprising most of the tree peony cultivars distributed throughout temperate regions in the world (Li et al., 2011). In addition to the ornamental use, the dried root bark of *P. suffruticosa* has been used in Chinese medicine for thousands of years for cardiovascular, extravasated blood, stagnated blood, and female genital diseases (Fu, Yang, Tsai, & Hsieh, 2012).

With long-term domestication and cultivation as well as natural and artificial selection, there are currently about 2,100 tree peony cultivars worldwide, and China alone has more than 1,000 cultivars (Li et al., 2011). The origin of these cultivars and relationships among them has attracted much attention, but remains unclear due to lack of detailed records of the complicated crossbreeding between wild species and cultivars during the long domestication history (Haw, 2001; Zhou et al., 2014). The numerous different flower colors and shapes of tree peony cultivars represent high genetic diversity and have been used for cultivars' classification in early studies (Zhou, Zhang, & Zhao, 2007). But based on molecular markers, the genetic groups of cultivars were not necessarily related to flower colors (Guo, Hou, & Zhang, 2009), while different provenances might be the more important factor contributing to the genetic differences (Liu & Lu, 2009; Yuan, Cheng, & Zhou, 2011). The high genetic diversity, along with wide geographic distribution, has made tree peony a fascinating model for studying the mechanisms of diversification and adaptation in plants.

Tree peony has 5 pairs of chromosomes ( $2n = 10$ ) (Cheng, 2007). The first high-density genetic map of tree peony was constructed using genotyping by specific-locus amplified fragment (SLAF) sequencing (Cai, Cheng, Wu, Zhong, & Liu, 2015). It contained 1,189 SLAF markers, spanning 920.699 cM with an average distance of 0.774 cM between adjacent markers. The genetic information of *P. suffruticosa* available to date includes three linkage maps (Cai et al., 2015; Guo et al., 2017; Zhang et al., 2019), 2,415 expressed sequence tags (ESTs) deposited to the NCBI database, and six RNA-seq datasets. Although analysis of the expressed sequences had contributed a lot to our understanding of the mechanisms of flower bud development (Shu et al., 2009), reblooming (Zhou, Cheng, Wang, Zhong, & He, 2013), prolonging vase life of cut flowers (Zhang et al., 2014), and different color formation in petals (Zhang, Cheng, Ya, Xu, & Han, 2015a) or leaves (Luo, Shi, Niu, & Zhang,

2017), our comprehensive and in-depth understanding of the genetic basis underlying the numerous flower morphological differences, oil production, and medicinal use is still limited. The availability of a reference genome sequence of *P. suffruticosa* would be helpful for the integration of multi-omics data across studies to enable more in-depth research into the biology and genetics of tree peony. Furthermore, a fully annotated genome of *P. suffruticosa* would serve as a foundation for cloning of important horticultural traits-related genes, identification of new varieties, and conservation of endangered varieties, as well as to promote more efficient breeding of tree peony.

In the present study, we report a de novo assembly and annotation of the *P. suffruticosa* genome, with an estimated genome size of ~13.66–15.76 Gb using PacBio's Single Molecule, Real-Time Technology (SMRT). Furthermore, we analyzed its phylogenetic relationship with closely related plants based on the genome sequences and reported a comprehensive analyses of the MADS-box gene family in this tree peony cultivar.

## 2 | MATERIAL AND METHODS

### 2.1 | Plant materials and sequencing

An individual of *P. suffruticosa* “Luo shen xiao chun” (Figure 1) grown in the peony resource spectrum of Luoyang Academy of Agriculture



**FIGURE 1** A flowering plant of *P. suffruticosa* “Luo shen xiao chun.”

and Forestry Sciences (N34°39' latitude, E112°27' longitude, Luoyang, China) was selected and the voucher specimen was deposited in the Herbarium of China National GeneBank with a code number "HCNGB\_00009295". The genomic DNA was isolated from the leaves with a standard CTAB extraction method (Murray & Thompson, 1980). A 20kb library was constructed as described previously (Pendleton et al., 2015). Approximately 20 µg of high-quality genomic DNA was sheared to ~20 kb targeted size and assessed with an Agilent 2,100 Bioanalyzer. Shearing of genomic DNA was followed by damage repair and end repair, blunt-end adaptor ligation, and size selection with a Blue Pippin system (Sage Science). A total of 114 SMRT cells and 177 SMRT cells were sequenced on PacBio RS II system and PacBio Sequel system, respectively. In total, 96.1 million subreads (894 Gb) were generated with an N50 of 14.5 kb and a mean length of 9.3kb. Furthermore, one paired-end library was constructed according to the standard protocol provided by BGI (BGI-Shenzhen) and sequenced on the BGISEQ-500 platform (Goodwin, McPherson, & McCombie, 2016), with a read length of 100 bp, generating a total of 673 Gb clean data.

Total RNA was extracted from the root, stem, shoot, leaf, flower, and flower bud tissues collected from the same individual using a rapid CTAB-based method as described previously (Gambino, Perrone, & Gribaudo, 2008). Paired-end libraries were constructed using standard protocol provided by BGI (BGI-Shenzhen) and then sequenced on the BGISEQ-500 platform, with a read length of 100 bp. In total, 45.71 Gb raw data were obtained and after filtering by SOAPnuke (Version 1.5.6) (<https://github.com/BGI-flexlab/SOAPnuke>), there were 6.85 ~ 8.46 Gb clean data for each sample (Supplementary Table S1).

## 2.2 | Estimation of *P. suffruticosa* genome size

A total of 520 Gb high-quality clean reads obtained from the BGISEQ-500 platform were subjected to 17, 19, 21, and 23 kmer frequency distribution analyses using Jellyfish (Marcais & Kingsford, 2011). The frequency graph (Supplementary Figure S1) was drawn and the *P. suffruticosa* genome size was calculated using the formula: genome size = kmer\_Number/Peak\_Depth. For 17 kmer, the total number of kmers was 519,130,610, and 870, and the peak depth was 38. The *P. suffruticosa* genome size was estimated to be 13.66 Gb, and the data used in 17 kmer analysis was about 46.53× coverage of the genome. In 19, 21, and 23 kmer analyses, the genome size was estimated to be 14.35, 14.77, and 15.76 Gb, respectively (Supplementary Table S2).

## 2.3 | Genome assembly and completeness assessment

Falcon v1.8.7 (Chin et al., 2016), a diploid-aware long-read assembler, was employed to assemble the PacBio subreads in this study. Error correction was first applied to the subreads using parameter "length\_cutoff = 13,000, pa\_HPCdaligner\_option = -v -B286

-t12 -w8 -M24 -e.75 -k18 -h380 -l2800 -s1000 -T4," and a total of 288 Gb corrected data were achieved. Then, these corrected reads were used to assemble the genome with parameter "length\_cutoff\_pr = 9,000, ovlp\_HPCdaligner\_option = -v -B180 -t12 -k18 -h180 -e.95 -l2200 -s1000, overlap\_filtering\_setting = --max\_diff 40 --max\_cov 60 --min\_cov 1." As a result, 11.8 Gb assembly with the contig N50 length of 76.7 kb was generated (Supplementary Table S4). Since the assembling of highly repetitive genome is sensitive to program parameters in FALCON pipeline, we tuned the parameter values of length\_cutoff\_pr and overlap\_filtering\_setting to explore the alternative assemblies (Supplementary Table S3). Finally, we obtained seven different assembly versions (Supplementary Table S4). Among them, the N50 length ranged from 48.4 kb in version 7 to 76.7kb in version 1, and the assembly size ranged from 11.5 Gb in version 2 to 13.8 Gb in version 6. In an overall view, the assembling result showed that no assembly was undoubtedly better than another one. To choose the most suitable genome assembly for functional genomic studies, we further evaluated the completeness of the seven assemblies by comparing them against a set of 1,440 conserved plant genes in BUSCO embryophyta\_odb9 dataset using BUSCO v2.0 (Simao, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015) pipeline. The completeness score ranged from 57.5% in version 3 to 61.2% in version 6 (Supplementary Table S5). We observed that the completeness score was not necessarily accordant with the contiguity (contig N50) among the seven assembly versions. Although version 1 has the highest N50 value (76.7 Kb), it captured less genome sequence than other versions and has a relatively smaller completeness score (58.3%). In fact, the assembly version 6 and 7 contained more total sequence than version 1 to 5, which might be due to the set of a smaller value of parameter "length\_cutoff\_pr = 6,000." We chose the assembly version 6 for further improvement because it had the highest completeness score and contained the most total sequences. High-quality BGISEQ-500 reads were mapped to this assembly with BWA-MEM (Li, 2013) with default parameters, and high-quality mapped reads (MAQ >20) were further used to polish the assembly with Pilon (<https://github.com/broadinstitute/pilon/wiki>) with default parameters. Finally, the obtained assembly was named V\_final, which had a total length of 13.79 Gb with an N50 length of 49.94 Kb.

In addition, four publicly available RNA-seq datasets for tree peonies (NCBI Short Read Archive, accession number: SRX336125, SRX314813, SRX698348, and SRX2439581) were aligned to the final assembly version using BLAT (Kent, 2002) for further validation.

## 2.4 | Annotation

Tandem repeats were identified using Tandem Repeats Finder v4.07b (Benson, 1999). For the transposable element annotation, RepeatMasker v3.3.0 (Tarailo-Graovac & Chen, 2009) and RepeatProteinMasker v3.3.0 (Tarailo-Graovac & Chen, 2009) were used against Repbase 16.10 (Jurka et al., 2005) to identify known repeats in the *P. suffruticosa* genome. De novo repeat identification

was conducted using RepeatModeler v1.0.5 (Price, Jones, & Pevzner, 2005) and LTR\_FINDER v1.0.5 (Xu & Wang, 2007) programs, followed by RepeatMasker v3.3.0 to achieve the final results.

Gene models were predicted using a combination of de novo prediction, homology-based prediction, and transcriptome-based prediction. For de novo prediction, Augustus (Stanke, Steinkamp, Waack, & Morgenstern, 2004) analysis was conducted on the repeat masked genome, with *Vitis vinifera* as the reference. For homology-based prediction, protein sequences of *Glycine max* (Schmutz et al., 2010), *Solanum lycopersicum* (Tomato Genome, 2012), *Vitis vinifera* (Jaillon et al., 2007), *Prunus persica* (International Peach Genome I, 2013), and *Arabidopsis thaliana* (Initiative, 2000) were aligned against *P. suffruticosa* genome using tBLASTn v2.2.26 (E-value  $\leq 1.0e-05$ ). Gene structure was predicted using GeneWise (Birney, Clamp, & Durbin, 2004). For transcriptome-based prediction, clean RNA-seq reads generated in this study were mapped to the assembly using TopHat v2.1.0 (Trapnell, Pachter, & Salzberg, 2009) and assembled into transcripts using Cufflinks (Trapnell et al., 2012), then open reading frames were predicted with hidden Markov model (HMM)-based training parameters. Results derived from the above methods were integrated by EVM (Haas et al., 2008) to produce a consensus gene set.

The predicted gene models were functionally annotated by aligning their protein sequences against the KEGG (Kanehisa & Goto, 2000), GO (Ashburner et al., 2000), SwissProt (Bairoch & Apweiler, 2000), TrEMBL, and NR protein databases with BLASTP (E-value  $\leq 1.0e-05$ ). Protein motifs and domains were identified by comparing the sequences against various domain databases, including PFAM, PRINTS, PANTHER, ProDom, PROSITE, and SMART using InterProScan v5.21–60.0 (Quevillon et al., 2005). For ncRNA annotation, tRNA genes were identified by tRNAscan-SE V1.23 (Lowe & Eddy, 1997). rRNA genes were identified by aligning the rRNA sequences from closely related species (*V. vinifera* and *K. fedtschenkoi*) against the assembly using BLASTN (E-value  $\leq 1.0e-05$ ). miRNAs and snRNAs were predicted by using INFERNAL (Nawrocki, Kolbe, & Eddy, 2009) software against the Rfam database (Griffiths-Jones et al., 2005).

## 2.5 | Gene family and phylogenetic analysis

For gene family analysis, OrthoMCL (Li, Stoeckert, & Roos, 2003) was used to construct orthologous gene families on all the protein-coding genes of *P. suffruticosa* and 7 sequenced plant species (*O. sativa* (Ouyang et al., 2007), *S. lycopersicum*, *C. roseus* (Kellner et al., 2015), *G. max*, *P. persica*, and *V. vinifera*, *Kalanchoë fedtschenkoi* (Yang et al., 2017)). Before OrthoMCL, BLASTP was performed to find similar matches from different species with an E-value cutoff of  $1.0e-05$ . The number of gene families in each species was calculated based on the composition of the OrthoMCL clusters. Genes that were single copy in an OrthoMCL cluster for all species analyzed were selected to construct phylogenetic trees using two methods. For the concatenation-based method, the protein sequences were

aligned using PRANK software version 170,427 (<http://wasabiapp.org/software/prank/>) and was then trimmed using Phyutility version 2.2.6 (Smith & Dunn, 2008) to remove poorly aligned regions with more than 30% missing data. The alignments were concatenated into one supermatrix file, which was used to reconstruct maximum-likelihood (ML) phylogenetic tree using IQ-TREE version 1.5.5 (Nguyen, Schmidt, Haeseler, & Minh, 2015) with automatic model selection and 1,000 bootstrap replicates. For coalescence-based method, individual ML gene trees were reconstructed from the CDS alignments using RAXML version 8.2.11 (Stamatakis, 2014) with GTRGAMMA model and 500 bootstrap replicates. The gene trees were used to reconstruct a species tree using ASTRAL v5.5.9 (Mirarab et al., 2014), with 1,000 bootstraps.

Species divergence times were estimated using MCMCTREE in PAML version 4.9 (Yang, 2007), based on the coalescent phylogenetic tree reconstructed from 511 single copy genes using ASTRAL. The ML estimates of branch lengths were obtained using CODEML programs in PAML under the JONES + gamma substitution models with the gamma priors set at 0.5. Two priors, the overall substitution rate (rgene gamma) and rate-drift parameter (sigma2 gamma), were set at G (1, 4.3) and G (1, 4.5). The correlated rates were used to specify the prior of rates among internal nodes (clock = 3 in MCMCTREE). The parameters of the birth-death process for tree generation with species sampling were fixed at BDparas = 1 1 0. A loose maximum bound for the root was set at <10.0 (=1,000 Ma). Markov chain Monte Carlo (MCMC) approximation with a burn-in period of 5,000,000 cycles was obtained, and every 5,000 cycles was taken to create a total of 10,000 samples. To diagnose possible failure of the Markov chains to converge to their stationary distribution, two replicate MCMC runs were performed with two different random seeds for each analysis. The stationarity of the chains and convergence of two runs were monitored by Tracer version 1.7 (Rambaut, Drummond, Xie, Baele, & Suchard, 2018) (<https://github.com/beast-dev/tracer/releases/tag/v1.7.1>). Divergence time estimates in TimeTree (Hedges, Marin, Suleski, Paymer, & Kumar, 2015) database were used for selecting the calibration priors. The lower and upper calibration values were chosen as 77–91, 82–116, and 110–124 for the most recent common ancestor (MRCA) of *C. roseus* and *S. lycopersicum*, *P. persica* and *G. max*, and eudicots, respectively. CAFÉ (De Bie, Cristianini, Demuth, & Hahn, 2006) was used to predict the expansion and contraction of gene family numbers based on the phylogenetic tree and gene family statistics.

## 2.6 | Identification of MADS-box genes from *P. suffruticosa* genome assembly and de novo transcriptome assembly

We identified putative MADS-box genes in *P. suffruticosa* genome using two methods. First, the 107 *Arabidopsis* MADS-box Protein sequences were used as query for BLASTP searches against the predicted *P. suffruticosa* protein sequences with an E-value cutoff of  $1.0e-05$ . Then, HMMER (Finn, Clements, & Eddy, 2011) searches were performed in *P.*

*suffruticosa* protein sequences using the hidden Markov model (HMM) profiles of MADS-box domain (PF00319) from the Pfam database (<http://pfam.janelia.org>). All putative MADS-box protein sequences obtained by the two methods were manually inspected for removing redundant gene sequences and confirming the existence of MADS-box domain according to their InterPro annotation. Meanwhile, if a sequence contained K domain in InterPro analysis, we classified it as type II MADS-box gene, otherwise as type I gene.

The transcriptome data produced in this study were used to search for potentially missing MADS-box genes. The RNA-seq reads of the six tissues were filtered using Trimmomatic v 0.38 (Bolger, Lohse, & Usadel, 2014) with parameters "HEADCROP: 15 LEADING: 20 TRAILING: 20 SLIDINGWINDOW: 5:20 MINLEN: 50 AVGQUAL: 20." Then, the clean reads were assembled using Trinity v2.4.09 (Grabherr et al., 2011) with a minimum contig length setting to 150 bp for each sample. We identified putative coding sequences (CDSs) within each transcript with TransDecoder (<https://transdecoder.github.io/>) using default parameter settings. We merged the CDSs of all samples and removed redundant sequences using CD-HIT-EST v4.6 (Li, 2006) with parameters "-c 0.8 -r 0." The putative MADS-box genes were identified from these CDSs using the same methods described above.

To further classify these genes into subfamilies, two individual phylogenetic trees for type I and type II genes were constructed using MADS-box protein sequences from *P. suffruticosa* and *Arabidopsis*. Multiple sequence alignment was performed using the Clustal X program (Larkin et al., 2007), and phylogenetic trees were then constructed using MEGA5 software (Tamura et al., 2011) with the neighbor-joining (NJ) method. Bootstrap values were calculated with 1,000 replicates to evaluate the support of the nodes.

## 2.7 | Expression analysis of *P. suffruticosa* MADS-box genes based on genome assembly and de novo transcriptome assembly

The expression profiles of the identified MADS-box genes in different tissues of *P. suffruticosa* were analyzed using the transcriptome data generated in this study. Because the two gene sets obtained

from the genome assembly and from de novo transcriptome assembly were largely different, we calculated the gene expression level for both separately. The filtered clean reads were mapped to the two gene sets using BOWTIE2 v2.2 (Langmead & Salzberg, 2012). The gene expression level was first quantified using RSEM program (Li & Dewey, 2011) and then was normalized by calculating the FPKM value for comparison between different samples.

## 3 | RESULTS

### 3.1 | Genome assembly and completeness assessment

Based on k-mer analyses, the *P. suffruticosa* genome was estimated to be ~13.66–15.76 Gb in size (Supplementary Figure S1). It is the largest genome in the sequenced dicots to date and presents a big challenge for genome sequencing and assembly. To assemble the *P. suffruticosa* genome, a total of 894 Gb third-generation long reads were generated using PacBio RS II system and PacBio Sequel system, representing ~67× coverage of the genome. Although PacBio reads have a relatively high error rate of ~15%, de novo assembly using these data was proved to be accurate enough with a deep coverage, typically >50× (Berlin et al., 2015). Falcon pipeline (Chin et al., 2016) was used to assemble the genome with significant parameter tuning, and seven different assembly versions were obtained. Completeness assessment of all the assemblies was performed using BUSCO (Simao et al., 2015) to choose the best assembly, followed by polishing with high-quality short reads. The final assembly version spanned 13.79 Gb in 499,810 contigs (N50 = 49.94 kb) (Table 1). Completeness assessment showed that 66.1% of the expected 1,440 plant conserved genes were detected as complete (Supplementary Table S5), which was comparable to that of recently sequenced mega-genomes (>10 Gb), such as 53% of sugar pine (Stevens et al., 2016) and 74% of *Ginkgo biloba* genome (Guan et al., 2016). Additionally, RNA sequence reads generated from six different tissues were mapped to our genome assembly by TopHat v2.1.0 (Trapnell et al., 2009) and the average mapping ratio was 73.2% (Supplementary Table S6). We also mapped four publicly available RNA-seq datasets to this assembly and found that the mapping ratio

**TABLE 1** Statistics of the final genome assembly

Category		Number	N50 (bp)	Size (bp)	Percentage of the assembly (%)
Contigs		499,810	49,937	13,793,297,086	100.00
Repetitive sequence				11,054,226,421	80.24
Transposable elements	LTR	—		6,874,219,419	49.90
	DNA			1,861,990,994	13.52
	LINE			1,931,149,253	14.02
	SINE			157,802,455	1.15
	Unknown			2,058,919,316	14.95
Annotated genes		35,687		6747/210/1192 <sup>a</sup>	

<sup>a</sup>Average mRNA length, exon length, and intron length, respectively.

was lower (0.04% ~ 68.41%, Supplementary Table S6), indicating high genetic diversity among different *P. suffruticosa* cultivars.

### 3.2 | Repeat analysis and gene prediction

In total, 11.05 Gb of repeat elements were identified, accounting for 80.24% of the 13.79-Gb genome assembly. Like other plant genomes, the long terminal repeat retrotransposons were the most abundant class of repetitive elements (49.9% of the assembled sequences), of which two superfamilies, Gypsy and Copia, account for 38.91% and 5.12% of the genome assembly, respectively (Supplementary Table S7). DNA class repeat elements represented 13.52% of the genome. Using RepeatMasker, we found that the sequence divergence rates of LTR-RTs were about 14 ~ 18% higher than other classes of transposon elements (Supplementary Figure S2), indicating that LTR-RTs played an important role in the genome evolution of tree peony.

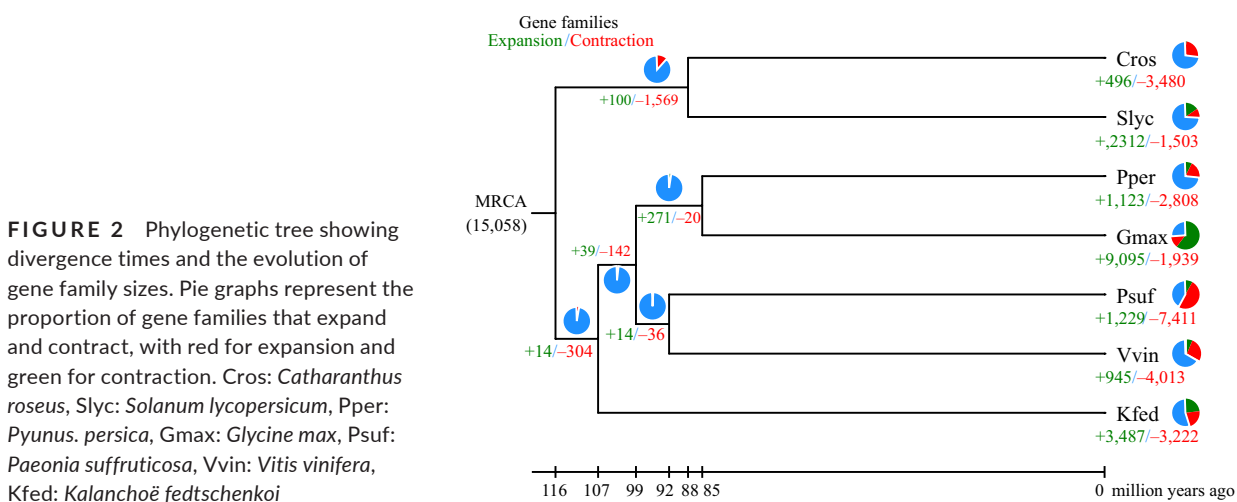
By integrating gene prediction results from ab initio, homology-based, and transcripts-based approaches, we predicted a nonredundant set of 35,687 gene models with an average gene length of 6,747 bp and an average coding sequence of 1,188 bp (Table 1; Supplementary Table S8). Of the 35,687 predicted genes, 51.37% were supported by either the identification of homologues in other species or RNA-seq data. The average lengths of gene, CDS, introns, and exons in *P. suffruticosa* were compared with selected six eudicots and were found to be similar to those reported for *V. vinifera* genome, indicating a relative close relation between them (Supplementary Table S9; Supplementary Figure S3). Functions were assigned to 34,854 (97.67%) genes, of which 32,258 (90.39%) had homology to proteins in SwissProt (Bairoch & Apweiler, 2000) and 31,885 (89.35%) had known protein domains in InterPro (Quevillon et al., 2005) (Supplementary Table S10). In addition to protein-coding genes, we predicted 1,215 tRNA, 1,510 rRNA, 960 microRNA (miRNA), and 1,055 small nuclear RNA (snRNA) genes in our assembly (Supplementary Table S11).

### 3.3 | Gene families and phylogenetic reconstruction

Using OrthoMCL (Li et al., 2003), the predicted 35,687 protein-coding genes in *P. suffruticosa* were assigned into 10,882 gene families consisting of 22,279 genes, while 13,408 genes were not organized into groups which might be mis-annotated or derived from lineage-specific expansion. Among these gene families, 1,794 were unique in *P. suffruticosa* compared with other seven plant species (Supplementary Table S12).

To infer phylogenetic relationships, a total of 511 single copy orthologs corresponding to the eight species were extracted from the clusters and were used to reconstruct phylogenetic trees with coalescence-based and concatenation-based method, respectively (Supplementary Figure S4a,b). The resulting species trees demonstrated a conflict on the placement of *K. fedtschenkoi*. In the species tree reconstructed from concatenated protein sequences, *K. fedtschenkoi* was placed as sister to the eudicots. In the coalescent species tree, *K. fedtschenkoi* was placed as sister to a clade of Vitales + Rosids, which was consistent with the APG IV tree (Chase et al., 2016). However, in both species trees reconstructed from the two methods, *P. suffruticosa* was placed as sister to Vitales, and they together formed a clade that was sister to Rosids. As the conflict between the concatenation-based tree and coalescence-based tree might indicate complicated evolutionary histories of genes in *K. fedtschenkoi*, we performed phylogenetic analyses excluding *K. fedtschenkoi* using the same two methods above. The resulting tree topologies (Supplementary Figure S4c,d) were congruent, supporting *P. suffruticosa* as sister to Vitales.

Based on the phylogenetic tree, *P. suffruticosa* was estimated to have separated from *V. vinifera* and *K. fedtschenkoi* approximately 92.3 and 98.9 Myr ago (Figure 2). The analysis of expansion and contraction of gene families between species using CAFÉ (De Bie et al., 2006) showed that 1,229 gene families were substantially expanded in *P. suffruticosa* and 7,411 gene families were contracted (Figure 2). The contraction is six times more than expansion, implying there are some missing genes due to the incomplete genome assembly or



indicating that *P. suffruticosa* has undergone large-scale gene loss events during the long domestication history.

### 3.4 | Identification of MADS-box genes from *P. suffruticosa* genome assembly and de novo transcriptome assembly

Using two methods of homology search of *Arabidopsis* MADS-box proteins and HMMER search of MADS-box domain profile, we identified 52 putative MADS-box genes in *P. suffruticosa* genome assembly, including 36 type I and 16 type II genes (Table S13). Based on the phylogenetic analysis, 36 type I genes were divided into M $\alpha$  and M $\gamma$  subgroups, which contained 19 and 17 members, respectively (Supplementary Figure S5). The M $\beta$  MADS-box genes are important in endosperm development (Masiero, Colombo, Grini, Schnittger, & Kater, 2011; Zhang et al., 2017), but are absent in our genome assembly. The number of type II MADS-box genes in *P. suffruticosa* genome assembly are significantly reduced, compared with that in other eudicots, such as 32 in *Prunus mume*, 64 in *Populus trichocarpa*, and 47 in *Arabidopsis thaliana*. The missing of representative member of SVP, ANR1, FLC, and AGL13 classes in the 16 type II MADS-box genes also indicated the incompleteness of the genome assembly.

In order to find potentially missing MADS-box genes in *P. suffruticosa* genome, we performed de novo transcriptome assembly and obtained 40,179 nonredundant CDS sequences, from which 8 type I and 24 type II MADS-box genes (Supplementary Table S14) were identified using the same methods described above. According to the phylogenetic tree, the missing MADS-box genes of M $\beta$ , SVP, ANR1, and AGL13 subfamilies in genome-based prediction were recovered from transcriptome-based prediction. No FLC orthologous gene was identified, indicating that this subfamily may have been lost in *P. suffruticosa*. In consideration of the possibility that the two MADS-box gene sets may have overlaps, we used the 32 transcriptome-derived MADS-box genes as queries to BLASTP against the 52 genome-derived MADS-box genes. As listed in Supplementary Table S15, there were only seven pairs of genes showed high protein identity (>90%). If considering each of the seven pairs of genes to be identical, we obtained a total of 77 MADS-box genes in *P. suffruticosa* in this study, including 44 type I and 33 type II genes. Of these type II genes, 18 were assigned to be ABCDE genes, including 5 A class genes, 5 B class genes, 2 C/D class genes, and 6 E class genes.

### 3.5 | Expression of MADS-box genes in different tissues

Transcriptome data showed that the expression of the type I MADS-box genes, except *PsuMADS5* and *TRINITYpsu2*, were all too weak to detect in the six different organs (Figure 3b). Since the six organs did not include the seed samples, we supposed that these type I genes might function in the seeds development as they do in

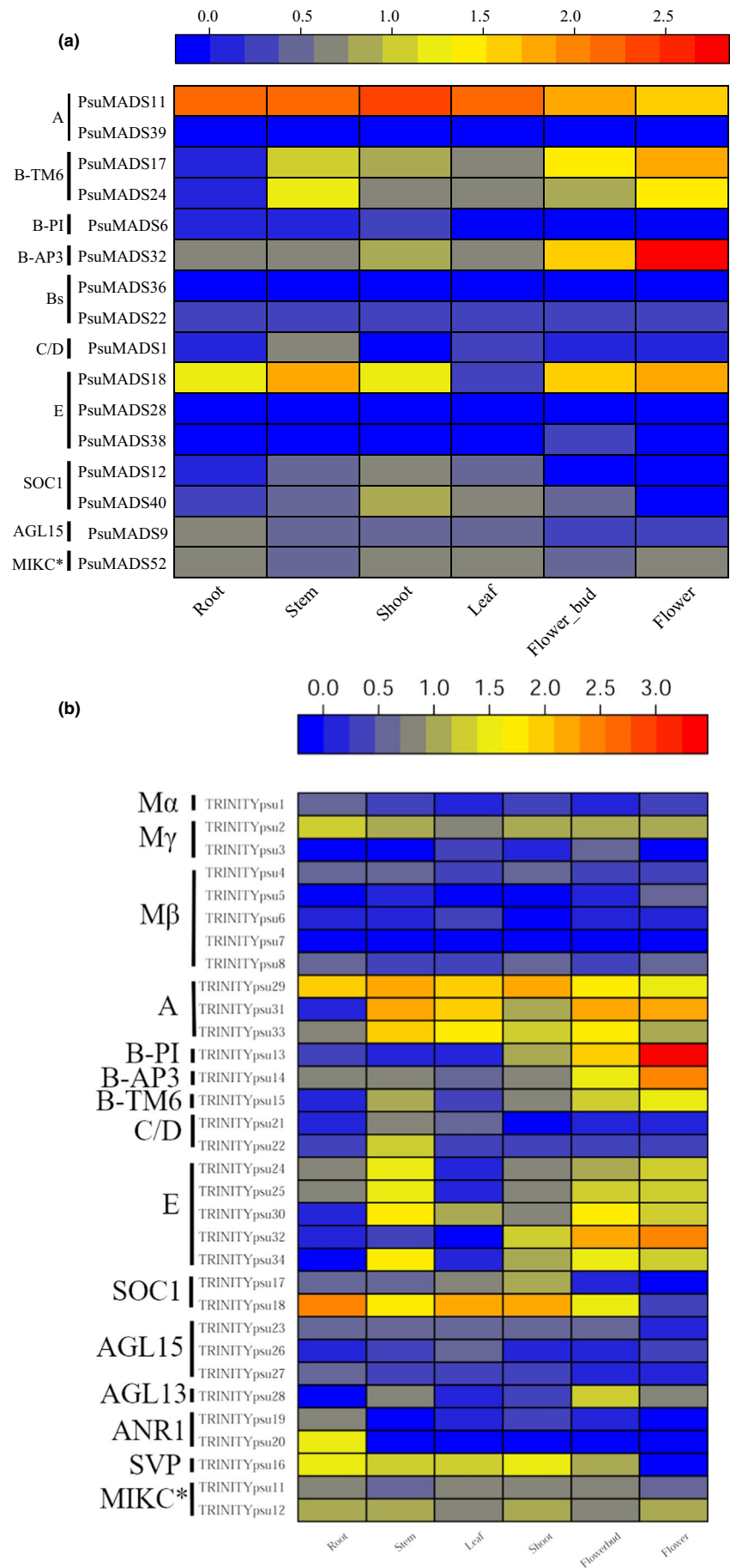
*Arabidopsis* (Bouyer et al., 2011; Kang, Steffen, Portereiko, Lloyd, & Drews, 2008).

In contrast to the weak expression of type I genes, most of the type II genes had a moderate or high expression in certain tissues (Figure 3a,b). The *SOC1* gene (*TRINITYpsu18*) was highly expressed in the vegetative tissues, in accordance with previous studies in tree peony (Zhang, Li, et al., 2015b). The *SVP* gene (*TRINITYpsu16*) was widely expressed in vegetative tissues and in flower bud, indicating that this gene may play multiple roles in tree peony development.

Most of the ABCDE genes exhibited the highest expressions in the flower bud and flower of *P. suffruticosa*, which conformed their important roles in flower development (Figure 3a,b). The A class genes (*PsuMADS11*, *TRINITYpsu29*, *TRINITYpsu31*, and *TRINITYpsu33*) were highly expressed in almost all six organs, implying they not only play the normal A class gene function in flower but also have multiple functions in other organs. Of the B class genes, *AP3* gene (*PsuMADS32* or *TRINITYpsu14*, they were nearly identical) and *B-PI* gene (*TRINITYpsu13*) had similar expression profile in flower and flower bud, implying they act as heterodimers on the formation of petals and stamens as they do in *Arabidopsis* and in other core eudicots (Riechmann, Krizek, & Meyerowitz, 1996; Wuest et al., 2012). The expression of another PI gene (*PsuMADS6*) could not be detected in neither flower organ, implying this gene had lost the B class gene function. The two B-TM6 genes (*PsuMADS17* and *PsuMADS24*, and *TRINITYpsu15* was nearly identical to *PsuMADS17*) had moderate to high expression in flower organs and stem, and weak expression in other three vegetative organs, which indicates that they may play multiple roles in tree peony. There were two C class genes (*PsuMADS1* and *TRINITYpsu22*, and *TRINITYpsu21* was nearly identical to *PsuMADS1*), which were homologous to *Arabidopsis* *AGAMOUS* (*AG*) gene (Supplementary Figure S6). Transcriptome data showed that the two C class genes were expressed at low level in flower organs. In the six E class genes, *PsuMADS18* (*TRINITYpsu24* and *TRINITYpsu25* were nearly identical to *PsuMADS18*), *TRINITYpsu30*, *TRINITYpsu32*, and *TRINITYpsu34* had moderate to high expression in flower organs, while *PsuMADS28* and *PsuMADS38* were barely expressed in any of the six organs.

## 4 | DISCUSSION

The genome assembly of *Paeonia suffruticosa* presented in this study represents the largest genome sequenced in the dicots, to date. It is a big challenge to assemble so large genome with high ratio of heterozygosity and repetition. The draft genome was fragmentary and was far from completed. The most possible reason for the poor assembly is that most of the PacBio subreads, with an N50 of 14.5 kb and a mean length of 9.3kb, cannot span the repetitive regions across the large chromosome. Another reason for the poor assembly may be the PacBio data were insufficient. Although ~67 $\times$  PacBio subreads were used, the error-corrected data used for next assembling step in FALCON pipeline was about 20 $\times$ . So, the key to improve the



**FIGURE 3** Expression profiles of MADS-box genes in six organs. (a) Expression profiles of type II MADS-box genes identified from the genome assembly. (b) Expression profiles of MADS-box genes identified from the de novo transcriptome assembly. Transcriptome sequencing was employed to investigate expression patterns of MADS-box genes. The colour scale shown at the top represents the normalized expression level ( $\log_{10}(\text{FPKM} + 1)$ ). Blue indicates low expression levels while red indicates high levels



genome assembly of *Paeonia suffruticosa* is to get more and longer sequencing reads, which should be taken into account in improvement of this genome assembly in future.

The draft genome of *P. suffruticosa* is the first genome sequenced for the Paeoniaceae, and the second sequenced species in Saxifragales. The coalescent phylogenetic tree (Supplementary Figure S4 (a)) placed the two sequenced species of Saxifragales into two different clades, implying paraphyly of the order of Saxifragales. According to this coalescent tree, the position of *P. suffruticosa* supports a relationship of ((Saxifragales and Vitales) and Rosids) and the position of *K. fedtschenkoi* supports a relationship of ((Rosids and Vitales) and Saxifragales). But the monophyly of Saxifragales has been strongly supported by molecular data in previous studies (Soltis et al., 2013; Soltis, Soltis, Endress, & Chase, 2005). Moreover, the proportion of gene trees that supported the tree topologies at each node showed that the positions of *P. suffruticosa* and *K. fedtschenkoi* in the coalescent tree were both weakly supported. So, we can reject neither of the possible position of Saxifragales in the phylogenetic tree. Although a "superrosid" clade of Saxifragales, Vitales, and Rosids is strongly supported, the relationships among these three groups have been debated for many years and different topologies were proposed by several studies based on plastid and nuclear genomes (Moore et al., 2011; Moore, Soltis, Bell, Burleigh, & Soltis, 2010; Soltis et al., 2011; Zeng et al., 2017). It is believed that the challenge of resolving the relationships among the three groups is ascribed to the rapid diversification of early eudicots (Magallon, Gomez-Acevedo, Sanchez-Reyes, & Hernandez-Hernandez, 2015; Moore et al., 2010) and the concomitant incomplete lineage sorting in phylogenetic tree reconstructing process (Degnan & Rosenberg, 2009). In previous studies, the split between Vitaceae and Saxifragales was dated to 112–101 mya based on dataset of plastid genes (Moore et al., 2010), and the split between Saxifragales and Rosids was dated to c. 112.4 mya based on dataset of nuclear genes (Zeng et al., 2017), indicating early and rapid diversification in superrosids. Thus, a better resolution among the three groups of superrosids needs further and more extensive genomic and taxon sampling, especially of Saxifragales species.

Tree peonies have an extensive history of domestication, during which the most important traits that have changed are the shapes and colors of the flowers, especially the number and colors of the petals. Most of the tree peony cultivars have various number of whorls of petals because of stamen petalody, but the inherent mechanism responsible for stamen or petal development in tree peony is still unclear. MADS-box transcription factors play important roles in plant development, especially in flower development. In the classical ABC(E) model of floral organ identities, A, B, C, and E all encode members of the MADS-box TF family, with A class genes specifying sepals, A + B + E specifying petals, B + C + E specifying stamens, and C + E specifying carpels (Litt & Kramer, 2010; Theissen & Saedler, 2001; Wellmer, Graciet, & Riechmann, 2014). According to the ABCDE model of flower development, B and C class MADS-box genes determined the identity of petal and stamen. In this study, we identified five nonredundant B class genes and two C class genes

from the combined dataset of genome assembly and transcriptome assembly. Except one B-PI gene, the other four B class genes were all highly expressed in flower. It was worth noting that the sequences of the two B-TM6 genes had been proved to be different among wild species and cultivars of tree peonies, which was explained to be related to stamen petalody and different flower shape formation in tree peonies (Shu et al., 2012). In the B class genes, TM6 are paralogs of AP3. The function of TM6 has been mostly studied in asterid model plant, such as petunia and tomato, in which TM6 orthologs are mainly expressed in stamens and carpels and may function redundantly in stamen identity with AP3 (Gemma, Pan, Emmanuel, Levy, & Irish, 2006; Rijpkema et al., 2006). So, it means TM6 may have part of C class gene function in spite of its belonging to B class. Meanwhile, the expression of C class genes in this study was very low in flower organs. In *Arabidopsis*, mutations in the AG gene result in the double flower phenotype, that is, the stamens are replaced by petals and carpels are replaced by a new flower (Yanofsky et al., 1990). The loss-of-function or restricted expression of the C function genes has shown to play a central role in the production of excessive numbers of petals in many different species, such as in *Thalictrum thalictroides* (Galimba et al., 2012), *Prunus lannesiana* (Liu, Zhang, Liu, Li, & Lu, 2013), *Camellia japonica* (Sun et al., 2014), and in rose (Dubois et al., 2010). During plant domestication, the causal mutations for convergent changes in key traits are likely to be located in particular genes (Lenser & Theissen, 2013). So, we suppose that the restriction of C class gene function may be responsible for the stamen petalody in tree peony cultivars. In addition, when the function of C class genes is restricted, the expression of TM6 genes can assure the normal development of carpels. It is reasonable to infer that the combined activity of AP3/PI and TM6 genes determines the formation of petal and stamen as well as the conversion between them when the C class gene function is restricted.

## 5 | CONCLUSION

This study presents the first genome in the family Paeoniaceae and the largest eudicots genome sequenced to date. This genome is also an important addition to Saxifragales genomic resources, which will facilitate the research of the phylogeny of this highly diverse clade. By integrating this genome with transcriptome data, we have demonstrated the use of this genome to explore the molecular mechanism underlying the flower development specified in this ornamental plant and suggested a modified BC model in the formation of petal and stamen. It can be expected that this genome will aid in deciphering the formation of specific and important traits in tree peony, such as various flower colors, oil accumulation in seeds, biosynthesis pathways of pharmacologically active metabolites, and adaption under domestication.

## ACKNOWLEDGMENTS

We are grateful to all participants of the Agriculture Department at BGI. We also thank the experts at the CNGB (China National

GeneBank, Shenzhen, China) for their kind help in voucher specimen deposition. The work was supported by Luoyang Municipal Government of China and Shenzhen Municipal Government of China (NO. JCYJ20160331150844452 and NO. JCYJ20150831201123287). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## AUTHORS' CONTRIBUTIONS

Canjun Zhang, Jiangtao Huang, Yang Dong, and Gengyun Zhang conceived and designed the project. Shuzuo Lv, Zhanying Wang, Erqiang Wang, and Xiaogai Hou collected the plant materials and performed the experiments. Bing Yang, Shu Cheng, Gang Huang, Jie Wang, Shiming Li, Xin Jin, Lei Lan, Kang Yu, Ning Li, and Xuemei Ni analyzed the data. Shu Cheng, Shuzuo Lv, Gengyun Zhang, and Canjun Zhang wrote the paper. All authors read and consented to the final version of the manuscript.

## DATA AVAILABILITY STATEMENT

All raw sequencing data, genome assembly sequences and gene annotations are available at the China National GeneBank (CNGB) Nucleotide Sequence Archive (CNSA) under BioProject accession code CNP0000281. The genome sequencing reads and genome assembly are deposited in CNSA with the Biosample accession CNS0044072. The RNA-seq reads are deposited in CNSA with the BioSample accession codes CNS0044073–CNS0044078.

## ORCID

Shu Cheng  <https://orcid.org/0000-0001-7077-4165>

## REFERENCES

- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25, 25–29. <https://doi.org/10.1038/75556>
- Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, 28, 45–48. <https://doi.org/10.1093/nar/28.1.45>
- Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research*, 27, 573–580. <https://doi.org/10.1093/nar/27.2.573>
- Berlin, K., Koren, S., Chin, C.-S., Drake, J. P., Landolin, J. M., & Phillippy, A. M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology*, 33, 623–630. <https://doi.org/10.1038/nbt.3238>
- Birney, E., Clamp, M., & Durbin, R. (2004). GeneWise and Genomewise. *Genome Research*, 14, 988–995. <https://doi.org/10.1101/gr.1865504>
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30, 2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>
- Bouyer, D., Roudier, F., Heese, M., Andersen, E. D., Gey, D., Nowack, M. K., ... Schnittger, A. (2011). Polycomb repressive complex 2 controls the embryo-to-seedling phase transition. *PLoS Genetics*, 7, e1002014. <https://doi.org/10.1371/journal.pgen.1002014>
- Cai, C., Cheng, F. Y., Wu, J., Zhong, Y., & Liu, G. (2015). The First high-density genetic map construction in tree peony (*Paeonia Sect. Moutan*) using genotyping by specific-locus amplified fragment. *Sequencing. Plos One*, 10, e0128584.
- Chase, M. W., Christenhusz, M. J. M., Fay, M. F., Byng, J. W., Judd, W. S., Soltis, D. E., ... Stevens, P. F. (2016). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society*, 181, 1–20.
- Cheng, F. Y. (2007). Advances in the breeding of tree peonies and a cultivar system for the cultivar group. *International Journal for Plant Breeding*, 1, 89–104.
- Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., ... Schatz, M. C. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, 13, 1050–1054. <https://doi.org/10.1038/nmeth.4035>
- Christenhusz, M. J. M., & Byng, J. W. (2016). The number of known plants species in the world and its annual increase. *Phytotaxa*, 261, 201. <https://doi.org/10.11646/phytotaxa.261.3.1>
- De Bie, T., Cristianini, N., Demuth, J. P., & Hahn, M. W. (2006). CAFE: A computational tool for the study of gene family evolution. *Bioinformatics*, 22, 1269–1271. <https://doi.org/10.1093/bioinformatics/btl097>
- De Martino, G., Pan, I., Emmanuel, E., Levy, A., & Irish, V. F. (2006). Functional analyses of two tomato APETALA3 genes demonstrate diversification in their roles in regulating floral development. *The Plant Cell*, 18, 1833.
- Degnan, J. H., & Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, 24, 332–340. <https://doi.org/10.1016/j.tree.2009.01.009>
- Dubois, A., Raymond, O., Maene, M., Baudino, S., Langlade, N. B., Boltz, V., ... Bendahmane, M. (2010). Tinkering with the C-function: A molecular frame for the selection of double flowers in cultivated roses. *PLoS ONE*, 5, e9288. <https://doi.org/10.1371/journal.pone.0009288>
- Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, 39, W29–37. <https://doi.org/10.1093/nar/gkr367>
- Fu, P. K., Yang, C. Y., Tsai, T. H., & Hsieh, C. L. (2012). Moutan cortex radicles improves lipopolysaccharide-induced acute lung injury in rats through anti-inflammation. *Phytomedicine*, 19, 1206–1215. <https://doi.org/10.1016/j.phymed.2012.07.013>
- Galimba, K. D., Tolkin, T. R., Sullivan, A. M., Melzer, R., Theissen, G., & Di Stilio, V. S. (2012). Loss of deeply conserved C-class floral homeotic gene function and C- and E-class protein interaction in a double-flowered ranunculid mutant. *Proceedings of the National Academy of Sciences*, 109, 13478–13479. <https://doi.org/10.1073/pnas.1203686109>
- Gambino, G., Perrone, I., & Griboaldo, I. (2008). A Rapid and effective method for RNA extraction from different tissues of grapevine and other woody plants. *Phytochemical Analysis*, 19, 520–525. <https://doi.org/10.1002/pca.1078>
- Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17, 333–351. <https://doi.org/10.1038/nrg.2016.49>
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29, 644–652. <https://doi.org/10.1038/nbt.1883>
- Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R., & Bateman, A. (2005). Rfam: Annotating non-coding RNAs in complete genomes. *Nucleic Acids Research*, 33, D121–124.
- Guan, R., Zhao, Y., Zhang, H. E., Fan, G., Liu, X., Zhou, W., ... Chen, W. (2016). Draft genome of the living fossil Ginkgo biloba. *Gigascience*, 5, 49. <https://doi.org/10.1186/s13742-016-0154-1>
- Guo, D.-L., Hou, X.-G., & Zhang, J. (2009). Sequence-related amplified polymorphism analysis of tree peony (*Paeonia suffruticosa* Andrews) cultivars with different flower colours. *The Journal of Horticultural Science and Biotechnology*, 84, 131–136.
- Guo, Q. I., Guo, L.-L., Zhang, L., Zhang, L.-X., Ma, H.-L., Guo, D.-L., & Hou, X.-G. (2017). Construction of a genetic linkage map in tree peony

- (*Paeonia* Sect. Moutan) using simple sequence repeat (SSR) markers. *Scientia Horticulturae*, 219, 294–301. <https://doi.org/10.1016/j.scienta.2017.03.017>
- Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., ... Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EvidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology*, 9, R7. <https://doi.org/10.1186/gb-2008-9-1-r7>
- Haw, S. (2001). Tree peonies: A review of their history and taxonomy. *New Plantsman*, 8, 156–171.
- Hedges, S. B., Marin, J., Suleski, M., Paymer, M., & Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification. *Molecular Biology and Evolution*, 32, 835–845. <https://doi.org/10.1093/molbev/msv037>
- Initiative AG (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408, 796.
- International Peach Genome I, Verde, I., Abbott, A. G., Scalabrin, S., Jung, S., ... Shu, S. (2013). The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nature Genetics*, 45(5), 487–494.
- Jaillon, O., Aury, J. M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., ... Vezzi, A. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449, 463–467.
- Ji, L., Wang, Q., Teixeira da Silva, J. A., & Yu, X. N. (2012). The genetic diversity of *Paeonia* L. *Scientia Horticulturae*, 143, 62–74. <https://doi.org/10.1016/j.scienta.2012.06.011>
- Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*, 110, 462–467. <https://doi.org/10.1159/000084979>
- Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28, 27–30. <https://doi.org/10.1093/nar/28.1.27>
- Kang, I. H., Steffen, J. G., Portereiko, M. F., Lloyd, A., & Drews, G. N. (2008). The AGL62 MADS domain protein regulates cellularization during endosperm development in *Arabidopsis*. *The Plant Cell*, 20, 635–647.
- Kellner, F., Kim, J., Clavijo, B. J., Hamilton, J. P., Childs, K. L., Vaillancourt, B., ... O'Connor, S. E. (2015). Genome-guided investigation of plant natural product biosynthesis. *The Plant Journal*, 82, 680–692. <https://doi.org/10.1111/tpj.12827>
- Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. *Genome Research*, 12, 656–664. <https://doi.org/10.1101/gr.229202>
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9, 357. <https://doi.org/10.1038/nmeth.1923>
- Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., ... Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, 23, 2947–2948. <https://doi.org/10.1093/bioinformatics/btm404>
- Lenser, T., & Theissen, G. (2013). Molecular mechanisms involved in convergent crop domestication. *Trends in Plant Science*, 18, 704–714. <https://doi.org/10.1016/j.tplants.2013.08.007>
- Li, B., & Dewey, C. N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12, 323. <https://doi.org/10.1186/1471-2105-12-323>
- Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.
- Li, J., Zhang, X., & Zhao, X. (2011). *Tree peony of China* (p. 206). Beijing: Encyclopaedia of China Publishing House.
- Li, L., Stoeckert, C. J. Jr, & Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13, 2178–2189. <https://doi.org/10.1101/gr.1224503>
- Li, S.-S., Wang, L.-S., Shu, Q.-Y., Wu, J., Chen, L.-G., Shao, S., & Yin, D.-D. (2015a). Fatty acid composition of developing tree peony (*Paeonia* section Moutan DC.) seeds and transcriptome analysis during seed development. *BMC Genomics*, 16, 208. <https://doi.org/10.1186/s12864-015-1429-0>
- Li, S.-S., Yuan, R.-Y., Chen, L.-G., Wang, L.-S., Hao, X.-H., Wang, L.-J., ... Du, H. (2015b). Systematic qualitative and quantitative assessment of fatty acids in the seeds of 60 tree peony (*Paeonia* section Moutan DC.) cultivars by GC-MS. *Food Chemistry*, 173, 133–140. <https://doi.org/10.1016/j.foodchem.2014.10.017>
- Li, W. (2006). Fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22, 1658.
- Litt, A., & Kramer, E. M. (2010). The ABC model and the diversification of floral organ identity. *Seminars in Cell and Developmental Biology*, 21, 129–137. <https://doi.org/10.1016/j.semcdb.2009.11.019>
- Liu, P., & Lu, M. (2009). AFLP analysis of genetic diversity of 35 cultivars of *Paeonia*, suffruticosa with 7 different flower colors. *Journal of Henan University of Chinese Medicine*, 3, 30–32.
- Liu, Z., Zhang, D., Liu, D., Li, F., & Lu, H. (2013). Exon skipping of AGAMOUS homolog PrseAG in developing double flowers of *Prunus lannesiana* (Rosaceae). *Plant Cell Reports*, 32, 227–237. <https://doi.org/10.1007/s00299-012-1357-2>
- Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, 25, 955–964. <https://doi.org/10.1093/nar/25.5.955>
- Luo, J., Shi, Q., Niu, L., & Zhang, Y. (2017). Transcriptomic analysis of leaf in tree peony reveals differentially expressed pigments genes. *Molecules*, 22, 324. <https://doi.org/10.3390/molecules22020324>
- Magallon, S., Gomez-Acevedo, S., Sanchez-Reyes, L. L., & Hernandez-Hernandez, T. (2015). A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytologist*, 207, 437–453. <https://doi.org/10.1111/nph.13264>
- Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27, 764–770. <https://doi.org/10.1093/bioinformatics/btr011>
- Masiero, S., Colombo, L., Grini, P. E., Schnittger, A., & Kater, M. M. (2011). The emerging importance of type I MADS box transcription factors for plant reproduction. *The Plant Cell*, 23, 865–872. <https://doi.org/10.1105/tpc.110.081737>
- Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., & Warnow, T. (2014). ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30, i541–548. <https://doi.org/10.1093/bioinformatics/btu462>
- Moore, M. J., Hassan, N., Gitzendanner, M. A., Bruenn, R. A., Croley, M., Vandeventer, A., ... Soltis, D. E. (2011). Phylogenetic analysis of the plastid inverted repeat for 244 species: Insights into deeper-level angiosperm relationships from a long, slowly evolving sequence region. *International Journal of Plant Sciences*, 172, 541–558. <https://doi.org/10.1086/658923>
- Moore, M. J., Soltis, P. S., Bell, C. D., Burleigh, J. G., & Soltis, D. E. (2010). Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proceedings of the National Academy of Sciences*, 107, 4623–4628. <https://doi.org/10.1073/pnas.0907801107>
- Murray, M. G., & Thompson, W. F. (1980). Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Research*, 8, 4321–4325. <https://doi.org/10.1093/nar/8.19.4321>
- Nawrocki, E. P., Kolbe, D. L., & Eddy, S. R. (2009). Infernal 1.0: Inference of RNA alignments. *Bioinformatics*, 25, 1335–1337. <https://doi.org/10.1093/bioinformatics/btp157>
- Nguyen, L. T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32, 268–274. <https://doi.org/10.1093/molbev/msu300>

- Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., ... Buell, C. R. (2007). The TIGR Rice Genome Annotation Resource: Improvements and new features. *Nucleic Acids Research*, 35, D883–887. <https://doi.org/10.1093/nar/gkl976>
- Pendleton, M., Sebra, R., Pang, A. W. C., Ummat, A., Franzen, O., Rausch, T., ... Bashir, A. (2015). Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods*, 12, 780–786. <https://doi.org/10.1038/nmeth.3454>
- Price, A. L., Jones, N. C., & Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics*, 21(Suppl 1), i351–358. <https://doi.org/10.1093/bioinformatics/bti1018>
- Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., & Lopez, R. (2005). InterProScan: Protein domains identifier. *Nucleic Acids Research*, 33, W116–120. <https://doi.org/10.1093/nar/gki442>
- Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior summarization in bayesian phylogenetics using Tracer 1.7. *Systematic Biology*, 67, 901–904. <https://doi.org/10.1093/sysbio/syy032>
- Riechmann, J. L., Krizek, B. A., & Meyerowitz, E. M. (1996). Dimerization specificity of Arabidopsis MADS domain homeotic proteins APETALA1, APETALA3, PISTILLATA, and AGAMOUS. *Proc Natl Acad Sci U S A*, 93, 4793–4798. <https://doi.org/10.1073/pnas.93.10.4793>
- Rijkema, A. S., Royaert, S., Zethof, J., van der Weerden, G., Gerats, T., & Vandenbussche, M. (2006). Analysis of the Petunia TM6 MADS box gene reveals functional divergence within the DEF/AP3 lineage. *The Plant Cell*, 18, 1819–1832.
- Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., ... Jackson, S. A. (2010). Genome sequence of the palaeopolyploid soybean. *Nature*, 463, 178–183. <https://doi.org/10.1038/nature08670>
- Shu, Q., Wang, L., Wu, J., Du, H., Liu, Z., Ren, H., & Zhang, J. (2012). Analysis of the formation of flower shapes in wild species and cultivars of tree peony using the MADS-box subfamily gene. *Gene*, 493, 113–123. <https://doi.org/10.1016/j.gene.2011.11.008>
- Shu, Q. Y., Wischnitzki, E., Liu, Z. A., Ren, H. X., Han, X. Y., Hao, Q., ... Wang, L. S. (2009). Functional annotation of expressed sequence tags as a tool to understand the molecular mechanism controlling flower bud development in tree peony. *Physiologia Plantarum*, 135, 436–449. <https://doi.org/10.1111/j.1399-3054.2009.01206.x>
- Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31, 3210–3212. <https://doi.org/10.1093/bioinformatics/btv351>
- Smith, S. A., & Dunn, C. W. (2008). Phylutility: A phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics*, 24, 715–716. <https://doi.org/10.1093/bioinformatics/btm619>
- Soltis, D. E., Mort, M. E., Latvis, M., Mavrodiev, E. V., O'Meara, B. C., Soltis, P. S., ... Rubio de Casas, R. (2013). Phylogenetic relationships and character evolution analysis of Saxifragales using a supermatrix approach. *American Journal of Botany*, 100, 916–929. <https://doi.org/10.3732/ajb.1300044>
- Soltis, D. E., Smith, S. A., Cellinese, N., Wurdack, K. J., Tank, D. C., Brockington, S. F., ... Soltis, P. S. (2011). Angiosperm phylogeny: 17 genes, 640 taxa. *American Journal of Botany*, 98, 704–730. <https://doi.org/10.3732/ajb.1000404>
- Soltis, D., Soltis, P., Endress, P., & Chase, M. (2005). *Phylogeny and evolution of angiosperms*. Sinauer Associates. Sunderland: Inc.
- Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30, 1312–1313. <https://doi.org/10.1093/bioinformatics/btu033>
- Stanke, M., Steinkamp, R., Waack, S., & Morgenstern, B. (2004). AUGUSTUS: A web server for gene finding in eukaryotes. *Nucleic Acids Research*, 32, W309–312. <https://doi.org/10.1093/nar/gkh379>
- Stevens, K. A., Wegrzyn, J. L., Zimin, A., Puiu, D., Crepeau, M., Cardeno, C., ... Langley, C. H. (2016). Sequence of the sugar pine megagenome. *Genetics*, 204, 1613–1626. <https://doi.org/10.1534/genetics.116.193227>
- Sun, Y., Fan, Z., Li, X., Liu, Z., Li, J., & Yin, H. (2014). Distinct double flower varieties in *Camellia japonica* exhibit both expansion and contraction of C-class gene expression. *Bmc Plant Biology*, 14, 288. <https://doi.org/10.1186/s12870-014-0288-1>
- Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, 28, 2731–2739. <https://doi.org/10.1093/molbev/msr121>
- Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences.
- Theissen, G., & Saedler, H. (2001). Plant biology. Floral quartets. *Nature*, 409, 469–471. <https://doi.org/10.1038/35054172>
- Tomato Genome, C. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485, 635–641.
- Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, 25, 1105–1111. <https://doi.org/10.1093/bioinformatics/btp120>
- Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., ... Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7, 562–578. <https://doi.org/10.1038/nprot.2012.016>
- Wellmer, F., Graciet, E., & Riechmann, J. L. (2014). Specification of floral organs in Arabidopsis. *Journal of Experimental Botany*, 65, 1–9. <https://doi.org/10.1093/jxb/ert385>
- Wuest, S. E., O'Maoileidigh, D. S., Rae, L., Kwasniewska, K., Raganelli, A., Hanczaryk, K., ... Wellmer, F. (2012). Molecular basis for the specification of floral organs by APETALA3 and PISTILLATA. *Proc Natl Acad Sci U S A*, 109, 13452–13457. <https://doi.org/10.1073/pnas.1207075109>
- Xu, Z., & Wang, H. (2007). LTR\_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, 35, W265–268. <https://doi.org/10.1093/nar/gkm286>
- Yang, X., Hu, R., Yin, H., Jenkins, J., Shu, S., Tang, H., ... Heyduk, K. (2017). The Kalanchoe genome provides insights into convergent evolution and building blocks of crassulacean acid metabolism. *Nature Communications*, 8, 1899.
- Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24, 1586–1591. <https://doi.org/10.1093/molbev/msm088>
- Yanofsky, M. F., Ma, H., Bowman, J. L., Drews, G. N., Feldmann, K. A., & Meyerowitz, E. M. (1990). The protein encoded by the Arabidopsis homeotic gene agamous resembles transcription factors. *Nature*, 346, 35–39. <https://doi.org/10.1038/346035a0>
- Yuan, J.-H., Cheng, F.-Y., & Zhou, S.-L. (2011). The phylogeographic structure and conservation genetics of the endangered tree peony, *Paeonia rockii* (Paeoniaceae), inferred from chloroplast gene sequences. *Conservation Genetics*, 12, 1539–1549. <https://doi.org/10.1007/s10592-011-0251-8>
- Zeng, L., Zhang, N., Zhang, Q., Endress, P. K., Huang, J., & Ma, H. (2017). Resolution of deep eudicot phylogeny and their temporal diversification using nuclear genes from transcriptomic and genomic datasets. *New Phytologist*, 214, 1338. <https://doi.org/10.1111/nph.14503>
- Zhang, C., Wang, Y., Fu, J., Dong, L. I., Gao, S., & Du, D. (2014). Transcriptomic analysis of cut tree peony with glucose supply using the RNA-Seq technique. *Plant Cell Reports*, 33, 111–129. <https://doi.org/10.1007/s00299-013-1516-0>

- Zhang, G.-Q., Liu, K.-W., Li, Z., Lohaus, R., Hsiao, Y.-Y., Niu, S.-C., ... Liu, Z.-J. (2017). The *Apostasia* genome and the evolution of orchids. *Nature*, *549*, 379–383. <https://doi.org/10.1038/nature23897>
- Zhang, L., Guo, D., Guo, L., Guo, Q. I., Wang, H., & Hou, X. (2019). Construction of a high-density genetic map and QTLs mapping with GBS from the interspecific F1 population of *P. ostii* 'Fengdan Bai' and *P. suffruticosa* 'Xin Riyuejin'. *Scientia Horticulturae*, *246*, 190–200. <https://doi.org/10.1016/j.scienta.2018.10.039>
- Zhang, Y., Cheng, Y., Ya, H., Xu, S., & Han, J. (2015a). Transcriptome sequencing of purple petal spot region in tree peony reveals differentially expressed anthocyanin structural genes. *Frontiers in Plant Science*, *6*, 964. <https://doi.org/10.3389/fpls.2015.00964>
- Zhang, Y., Li, Y. E., Zhang, Y., Guan, S., Liu, C., Zheng, G., & Gai, S. (2015b). Isolation and characterization of a SOC1-Like gene from tree peony (*Paeonia suffruticosa*). *Plant Molecular Biology Reporter*, *33*, 855–866. <https://doi.org/10.1007/s11105-014-0800-7>
- Zhou, H., Cheng, F. Y., Wang, R., Zhong, Y., & He, C. (2013). Transcriptome comparison reveals key candidate genes responsible for the unusual reblooming trait in tree peonies. *PLoS ONE*, *8*, e79996. <https://doi.org/10.1371/journal.pone.0079996>
- Zhou, S. L., Zou, X. H., Zhou, Z. Q., Liu, J., Xu, C., Yu, J., ... Sang, T. (2014). Multiple species of wild tree peonies gave rise to the 'king of flowers' (p. 281). *Proc Biol Sci: Paeonia suffruticosa* Andrews.
- Zhou, X. W., Zhang, Y. X., & Zhao, G. D. (2007). Research progress of the relationship of tree peonies. *Journal of Anhui Agricultural Sciences*, *35*, 391–393.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Lv S, Cheng S, Wang Z, et al. Draft genome of the famous ornamental plant *Paeonia suffruticosa*. *Ecol Evol.* 2020;10:4518–4530. <https://doi.org/10.1002/ece3.5965>