DNA Research, 2022, 29(2), 1–9
https://doi.org/10.1093/dnares/dsac012
Advance Access Publication Date: 19 April 2022
Resource Article: Genomes Explored

Resource Article: Genomes Explored

Chromosomal-level genome assembly of the orchid tree *Bauhinia variegata* (Leguminosae; Cercidoideae) supports the allotetraploid origin hypothesis of *Bauhinia*

Yan Zhong[1†], Yong Chen[2†], Danjing Zheng[2], Jingyi Pang[1], Ying Liu[1], Shukai Luo[2], Shiyuan Meng[2], Lei Qian[2], Dan Wei[3], Seping Dai[2*], and Renchao Zhou[1*]

[1]State Key Laboratory of Biocontrol and Guangdong Provincial Key Laboratory of Plant Resources, School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China, [2]Guangzhou Institute of Forestry and Landscape Architecture, Guangzhou 510405, China, and [3]Guangdong Academy of Forestry, Guangdong Provincial Key Laboratory of Silviculture, Protection and Utilization, Guangzhou 510520, China

*To whom correspondence should be addressed. Tel. +86-20-36377485. Email: daiseping@126.com (S.D.); Tel. +86-20-84111164. zhrench@mail.sysu.edu.cn (R.Z.)

†The first two authors contributed equally to this work.

Received 22 January 2022; Editorial decision 13 April 2022; Accepted 14 April 2022

Abstract

Cercidoideae, one of the six subfamilies of Leguminosae, contains one genus *Cercis* with its chromosome number $2n=14$ and all other genera with $2n=28$. An allotetraploid origin hypothesis for the common ancestor of non-*Cercis* genera in this subfamily has been proposed; however, no chromosome-level genomes from Cercidoideae have been available to test this hypothesis. Here, we conducted a chromosome-level genome assembly of *Bauhinia variegata* to test this hypothesis. The assembled genome is 326.4 Mb with the scaffold N50 of 22.1 Mb and contains 37,996 protein-coding genes. The Ks distribution between gene pairs in the syntenic regions indicates two whole-genome duplications (WGDs): one is *B. variegata*-specific, and the other is shared among core eudicots. Although Ks between gene pairs generated by the recent WGD in *Bauhinia* is greater than that between *Bauhinia* and *Cercis*, the WGD was not detected in *Cercis*, which can be explained by an accelerated evolutionary rate in *Bauhinia* after divergence from *Cercis*. Ks distribution and phylogenetic analysis for gene pairs generated by the recent WGD in *Bauhinia* and their corresponding orthologs in *Cercis* support the allopolyploidy origin hypothesis of *Bauhinia*. The genome of *B. variegata* also provides a genomic resource for dissecting genetic basis of its ornamental traits.

Key words: *Bauhinia variegata*, genome assembly, whole-genome duplication, allopolyploidization, rapid evolution

1. Introduction

Leguminosae is an economically and agronomically important family, with six subfamilies (Papilionoideae, Caesalpinioideae, Detarioideae, Cercidoideae, Dialioideae and Duparquetioideae), ca. 770 genera and 20,000 species.[1,2] Some legumes are major sources of plant protein and micronutrients, and have been used as high-quality food and fodder.[3]

Many legumes show high horticultural value and have been cultivated throughout the world. Given their economical and agronomical significance, genome sequencing has been conducted for quite a few legumes, mainly from Papilionoideae, including *Glycine max*,[4] *Cajanus cajan*,[5] *Arachis duranensis*,[6] *Pisum sativum*,[7] *Lotus japonicus*[8] and *Medicago truncatula*.[9] In contrast, draft genome sequences are available for only two species (*Mimosa pudica* and *Chamaecrista fasciculata*) of Caesalpinioideae and one species (*Cercis canadensis*) of Cercidoideae.[10]

Whole-genome duplication (WGD) plays important roles in plant genome evolution and diversification.[11,12] A previous study[13] showed that a WGD occurred in the common ancestor of all papilionoids (i.e. Papilionoideae) and several independent WGDs near the base of Caesalpinioideae, Detarioideae and Cercidoideae. Cercidoideae is the earliest-diverging subfamily among the six subfamilies of Leguminosae.[1] In Cercidoideae, *Cercis* is the only genus that has the chromosome number of $n = 7$, identical with the ancestral chromosome number inferred for legumes,[14] and all other genera have their chromosome number of $n = 14$ (CCDB; http://ccdb.tau.ac.il/). Genomic analysis of *Cercis* and other species of this subfamily suggested the lack of a recent WGD in *Cercis* and an allotetraploid origin for the common ancestor of the rest of the subfamily was proposed.[14] However, no chromosome-level fully assembled genome from Cercidoideae has been available to test this hypothesis.

Bauhinia, the largest genus of the subfamily Cercidoideae, consists of ∼380 species distributed in the pantropical regions,[1] with many species exhibiting high ornamental value and being widely cultivated in tropical regions. *Bauhinia variegata*, also called the orchid tree, possesses diverse petal colours varying from white to deep purple and is especially attractive in horticulture.

Here, we assembled the chromosomal-level genome of *B. variegata* using PacBio and Illumina sequencing, and Hi-C scaffolding technologies. Genome evaluation and annotation, phylogenomic analysis, gene family evolution and intra- and inter-genome synteny analysis were performed. We aimed to test the hypothesis of the allotetraploid origin of *Bauhinia* with the high-quality genome.

2. Materials and methods

2.1. Sampling and sequencing

Samples of an individual of *B. variegata* used for the whole-genome and transcriptome sequencing were obtained from Sun Yat-sen University campus, Guangzhou, China. Genomic DNA was extracted from the leaves. RNAs were isolated from four fresh tissues, i.e. flower, fruit, leaf and root. A DNA library with an insert size of 30 kb was constructed and then sequenced on the PacBio Sequel II System and 175.4 Gb reads were generated. To perform the genome survey, a short genome fragment library with an insert size of 350 bp was constructed and then sequenced on an Illumina NovaSeq platform, and 48.8 Gb paired-end reads of 150 bp were generated. Transcriptome sequencing was also conducted on the same Illumina NovaSeq platform and about 6 Gb sequence data were generated for each tissue.

For High-throughput Chromatin Conformation Capture (Hi-C), fresh leaves were cut into small pieces and infiltrated in 2% formaldehyde. Glycine was added to stop crosslinking. The tissue was ground to powder and nuclei isolation buffer was then added to obtain a nuclei suspension. Nuclei were digested with *HindIII* restriction endonuclease. DNA fragments of 150–300 bp were purified, and PCR amplification was performed after adapters were ligated to the Hi-C products. The PCR products were purified, and the Hi-C libraries were quantified by quantitative PCR for Illumina HiSeq X Ten sequencing. Finally, a total of 31.3 Gb paired-end reads of 150 bp were generated.

2.2. Genome size estimation

Genome survey analysis was performed using clean Illumina reads filtered by fastp 0.20.1[15] and FastUniq[16] with default parameters. K-mers were counted and k-mer count histogram was produced with Jellyfish v.2.3.0[17] for 48.8 Gb Illumina reads with k-mer length of 17. Genome size was estimated based on k-mer frequency distributions by GenomeScope 1.0 (http://qb.cshl.edu/genomescope/).

2.3. Genome assembly

The PacBio reads were corrected, trimmed and assembled into contigs using Canu v2.0[18,19] with the parameters correctedErrorRate = 0.035 and minReadLength = 2,000. The primary assembly was polished by referring to the PacBio reads and Illumina reads with NextPolish 1.2.0 with default parameters.[20] Finally, haplotigs and contig overlaps in the polished assembly were purged based on read depth using Purge_Dups (https://github.com/dfguan/purge_dups).

Hi-C unique reads were used to scaffold the PacBio assembly contigs using 3D-DNA pipeline. Hi-C datasets were first processed by Juicer.[21] Abnormal contact patterns in initially assembled contigs were corrected, partitioned, orientated and ordered, and finally anchored onto 14 pseudo-chromosomes using 3D-DNA.[22] We further manually adjusted the Hi-C scaffolding based on the chromatin contact matrix in Juicebox.[23]

2.4. Genome quality evaluation

The quality of the *B. variegata* genome was further evaluated based on eudicots_odb10 database (2326 BUSCOs) and fabales_odb10 database (5366 BUSCOs) using Benchmarking Universal Single-Copy Orthologs (BUSCO) programme[24] with default parameters. The same evaluation was also performed for the genomes of *C. canadensis*, *C. fasciculata*, *G. max* and *M. truncatula*.

2.5. Genome annotation

Known repeat sequences were identified by RepeatMasker v 4.1.1 (http://www.repeatmasker.org) with the Repbase library.[25] A *de novo* repeat library was constructed using RepeatModeler v 2.0.1.[26] RNA-seq data from four tissues were mapped to the genome by HISAT2,[27] merged by SAMtools,[28] and then transcripts were extracted by StringTie v 2.1.3[29] and coding regions in the transcripts were predicted by TransDecoder (https://github.com/TransDecoder/TransDecoder). The training result of RepeatModeler and the coding sequence from TransDecoder v 5.5.0 were supplied to EDTA[30] to identify repetitive sequences.

We predicted protein-coding genes using a combination of homologous-sequence search, *ab initio* gene prediction, and transcriptome-data comparison in an automatic genome annotation tool GETA v2.4.5 (https://github.com/chenlianfu/geta). Illumina RNA-seq reads from different tissues were used to assemble transcripts and predict genes using HISAT2[27] and TransDecoder. Protein sequences from Swiss-Prot plant database (https://www.uniprot.org/) and four legumes (*Arachis hypogaea*, *G. max*, *M. truncatula* and *Vigna unguiculata*) (Table 1) were combined for homology-based prediction with GeneWise (https://www.ebi.ac.uk/~birney/wise2/). *Ab initio* prediction was performed in Augustus v3.3.3,[31] trained with intron and exon information generated above. These prediction

Table 1. Sources of genomic and transcriptomic data of other species included in the study

Species	Sequence type	Source
Glycine max	Genomic	Phytozome
Medicago truncatula	Genomic	Phytozome
Lotus japonicus	Genomic	Phytozome
Vigna unguiculata	Genomic	Phytozome
Cercis canadensis	Genomic	GigaDB
Chamaecrista fasciculata	Genomic	GigaDB
Acacia pycnantha	Transcriptomic	http://www.onekp.com
Copaifera officinalis	Transcriptomic	http://www.onekp.com
Gleditsia triacanthos	Transcriptomic	http://www.onekp.com
Quillaja saponaria	Transcriptomic	http://www.onekp.com
Xanthocercis zambesiaca	Transcriptomic	http://www.onekp.com

results were integrated and then were searched against the Pfam database for screening to get the final gene prediction result. Functional annotation of genes was also performed by using InterProScan,[32] eggnog-mapper (http://eggnog-mapper.embl.de/), PANNZER2[33] and Mercator4 v3.0.[34] The functional annotation results were then integrated by an in-house script.

The density of genes, repeats, genes located in syntenic regions (see below) and GC content in 14 pseudo-chromosomes were calculated in a 100-kb sliding window with BEDTools v2.30.0[35] and were plotted with Circos v 0.69-8.[36]

2.6. Phylogenomic analysis

The longest protein or transcript data from nine legume species (*G. max*, *M. truncatula*, *L. japonicus* and *Xanthocercis zambesiaca* from Papilionoideae; *Acacia pycnantha*, *C. fasciculata* and *Gleditsia triacanthos* from Caesalpinioideae; *C. canadensis* from Cercidoideae and *Copaifera officinalis* from Detarioideae) and one outgroup (*Quillaja saponaria*) were downloaded (Table 1). All-against-all comparison was performed in OrthoFinder2[37] with default parameters based on protein sequences of the 11 species. For each ortholog, the protein sequences were aligned using PRANK,[38] and then converted into nucleotide sequence alignments using pal2nal.pl script.[39] All the sequence alignments were then concatenated into a supermatrix, and used for phylogenomic analyses. ModelTest-NG[40] with the Bayesian information criterion was employed for DNA substitution model selection, and RAxML-NG v 0.9.0[41] was used to construct a phylogenetic tree with 1,000 bootstrap replicates. The divergence time in the ML tree was estimated by mcmctree programme in the PAML package[42] with two soft calibration points between *Q. Saponaria* and *G. max*, *A. pycnantha* and *G. max* from TimeTree (http://www.timetree.org).

2.7. Gene family expansion and contraction analysis

The orthogroup information identified above and the phylogenetic tree constructed above were used to infer gene family expansion and contraction in CAFE5.[43] Gene families with >100 gene copies were filtered by the script clade_and_size_filter.py. Root frequency distribution was designated as the Poisson distribution, and the Gamma model was set with five gamma rate categories. Gene families with an accelerated rate of expansion and contraction were determined with a threshold conditional *P*-value ($P < 0.05$). The numbers of expanded and contracted gene families were labelled in the phylogenetic tree.

KEGG pathway enrichment analysis was conducted using KOBAS 3.0[44] with the parameters of top cluster = 5 and edge weight = 0.35, and statistical significance was tested by Fisher's exact test in combination with the False Discovery Rate correction.

2.8. Identification of WGD

All-versus-all alignment of the protein sequences of *B. variegata* was constructed using the Blastp algorithm.[45] To detect the signature of WGD, the programme MCScanX[46] with default parameters was used to define syntenic blocks. For each gene pair in the syntenic blocks, Ks value was calculated using KaKs_Calculator 2.0[47] with the YN model and the distribution of Ks values of all gene pairs was plotted using R package ggplot2.[48] Intragenomic synteny was plotted with Circos v 0.69-8.[36]

Meanwhile, inter-genomic syntenic blocks between *B. variegata* and *C. canadensis* were searched, and the Ks values between syntenic gene pairs were calculated as stated above. To show the genomic synteny between the two species, syntenic regions between the 14 chromosomes of *B. variegata* and 11 longest contigs of *C. canadensis* were identified and plotted with MCScan pipeline.[49]

2.9. Testing the allopolyploidy origin hypothesis

To test the allopolyploidy origin hypothesis of *Bauhinia*, gene pairs with the Ks range of 25% greater and lower than the Ks peak value for the *B. variegata*-specific WGD were extracted, and each of the extracted gene pairs was randomly assigned to two groups (B1 and B2). Orthologs were identified respectively with OrthoFinder2 for each of the two groups and two closely related species (*C. canadensis* and *C. fasciculata*). Shared single copy orthologs for the two groups were used for further analyses.

Amino acid sequences of each single-copy ortholog (homeologous B1 and B2 for *B. variegate* and their corresponding ortholog in *C. canadensis*) were aligned with MAFFT v 6.8,[50] and then converted into nucleotide sequences using ParaAT.[51] Ks values between B1 and B2, B1 and *C. canadensis*, and B2 and *C. canadensis* for each gene were calculated using the same method mentioned above. Ks distribution was plotted by R package ggplot2.

For phylogenetic analysis among B1, B2 and *C. canadensis*, one maximum likelihood tree was constructed with RAxML-NG[41] based on coding region sequences of each single copy ortholog, with *C. fasciculata* as an outgroup. The number of each tree topology was counted.

3. Results and discussion

3.1. Genome assembly and assembly quality assessment

We generated 175.4 Gb PacBio and 48.8 Gb Illumina reads from an individual of *B. variegata* and used them to assemble its genome. Genome survey of Illumina reads indicated that *B. variegata* has a genome size of 327.00 Mb (Fig. 1A). We obtained a genome assembly of 411 contigs with a total size of 326.4 Mb (Table 2), representing 99.8% of the estimated genome size. 92.2% (300.8 Mb) of sequences were anchored to the 14 pseudochromosomes based on the Hi-C data. The scaffold N50 and contig N50 are 22.09 Mb and 4.55 Mb, respectively. The overall GC content of the *B. variegata* genome is 35.0% (Table 2). This is the first chromosomal-level genome assembly for the subfamily Cercidoideae. *Bauhinia variegata* has the second smallest genome size among legumes with available genome

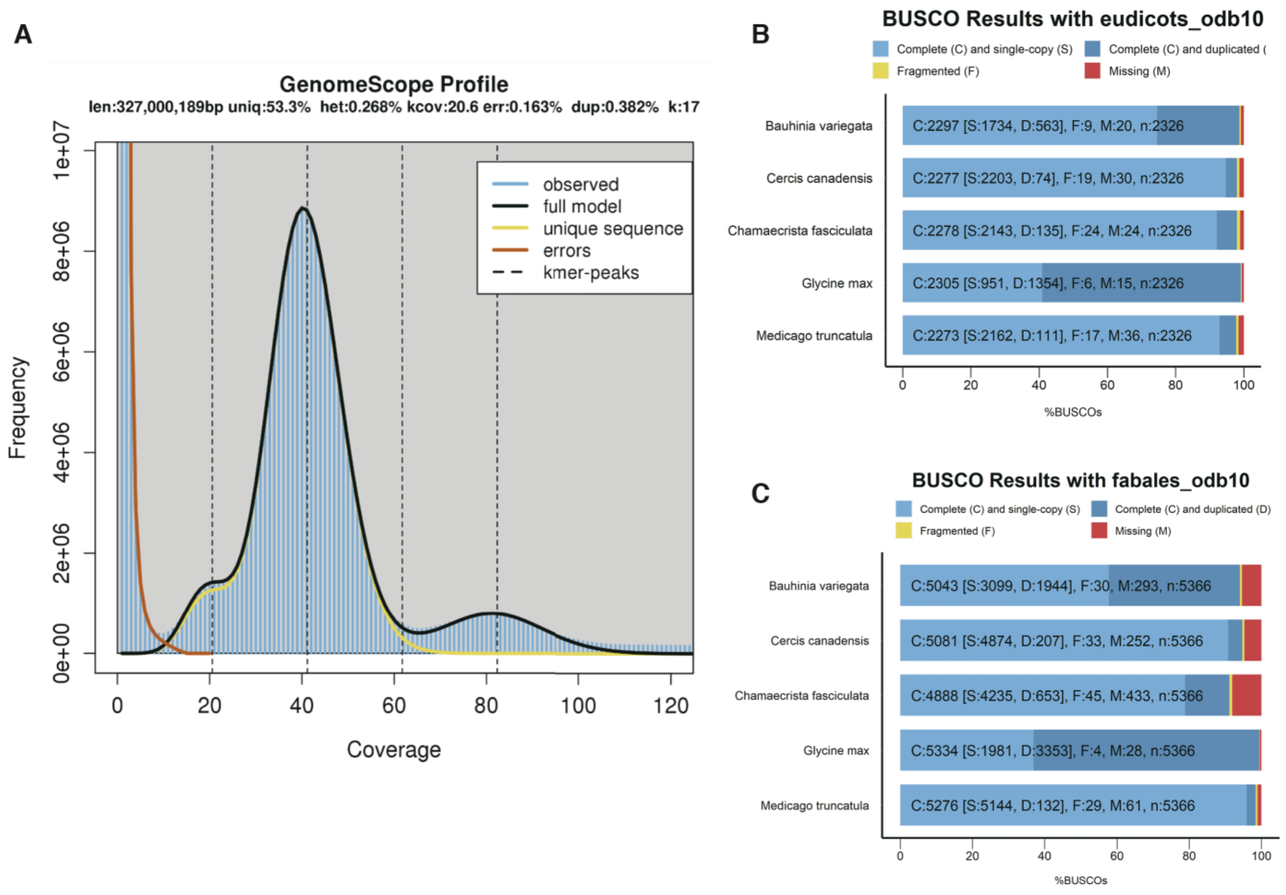

Figure 1. Genome size estimation and genome assembly assessment. (A) Genome survey of *Bauhinia variegata* with GenomeScope. (B) BUSCO assessment of the genome assemblies of five legumes with eudicots_odb10 dataset. (C) BUSCO assessment of the genome assemblies of five legumes with fabales_odb10 dataset.

Table 2. Statistics of the genome assembly for *Bauhinia variegata*

Assembly features	
Genome size (bp)	326,375,084
GC content	34.95%
Scaffolds number	411
Scaffold N50 (bp)	22,089,475
Scaffold L50	7
Contig N50 (bp)	4,549,988
Contig L50	21
Annotation features	
Number of predicted gene models	37,996
Mean of exon number per gene	5.4
Mean of exon length (bp)	297.5
Mean of intron length (bp)	382.5
Repeat content (% of the genome assembly)	27.22%
Functional annotation	
Total number of annotated genes	35,659
Number of genes annotated by InterProScan	35,189
Number of genes annotated by Eggnog	34,601
Number of genes annotated by Pannzer2	29,589
Number of genes annotated by Mercator4	26,311

N50: sequence length of the shortest contig/scaffold at 50% of the total genome length.

L50: the smallest number of contigs/scaffolds whose length sum makes up half of genome size.

size data (http://data.kew.org/cvalues/), only larger than *Leucaena macrophylla* (303 Mb).

The BUSCO analysis recovered 2,297 (98.7%) universal single copy genes of eudicots_odb10 dataset (2,326 genes) and 5,043 (94.0%) of fabales_odb10 (5,366 genes) in *B. variegata* (Fig. 1), indicating high completeness of the genome assembly. Comparative analysis among 10 legumes showed that *B. variegata* had the second highest proportion of duplicated complete BUSCOs (24.2% in eudicots_odb10 and 36.2% in fabales_odb10), only lower than soybean (58.2% and 62.5%, respectively), which has experienced two WGDs after the origin of legumes.[13] The high proportion of duplicated BUSCOs in *B. variegata* implies that there might be WGD(s) in this species (see below).

3.2. Genome annotation

Transposable elements took up 27.2% of the *B. variegata* genome (Table 2; Fig. 2c), including 8.6% LTR (4.2% Gypsy, 2.6% Copia and 1.9% others) and 12.0% TIR. Tandem repeat took up 0.64% of the genome. We identified 37,996 protein-coding genes in *B. variegata* based on *de novo* prediction, transcript evidence and homology with other known plant proteins (Table 2; Fig. 2b); 93.9% of the predicted genes were functionally annotated by at least one of the four databases (Table 2). The mean exon and intron sizes are 297.5 bp and 382.5 bp, respectively (Table 2).

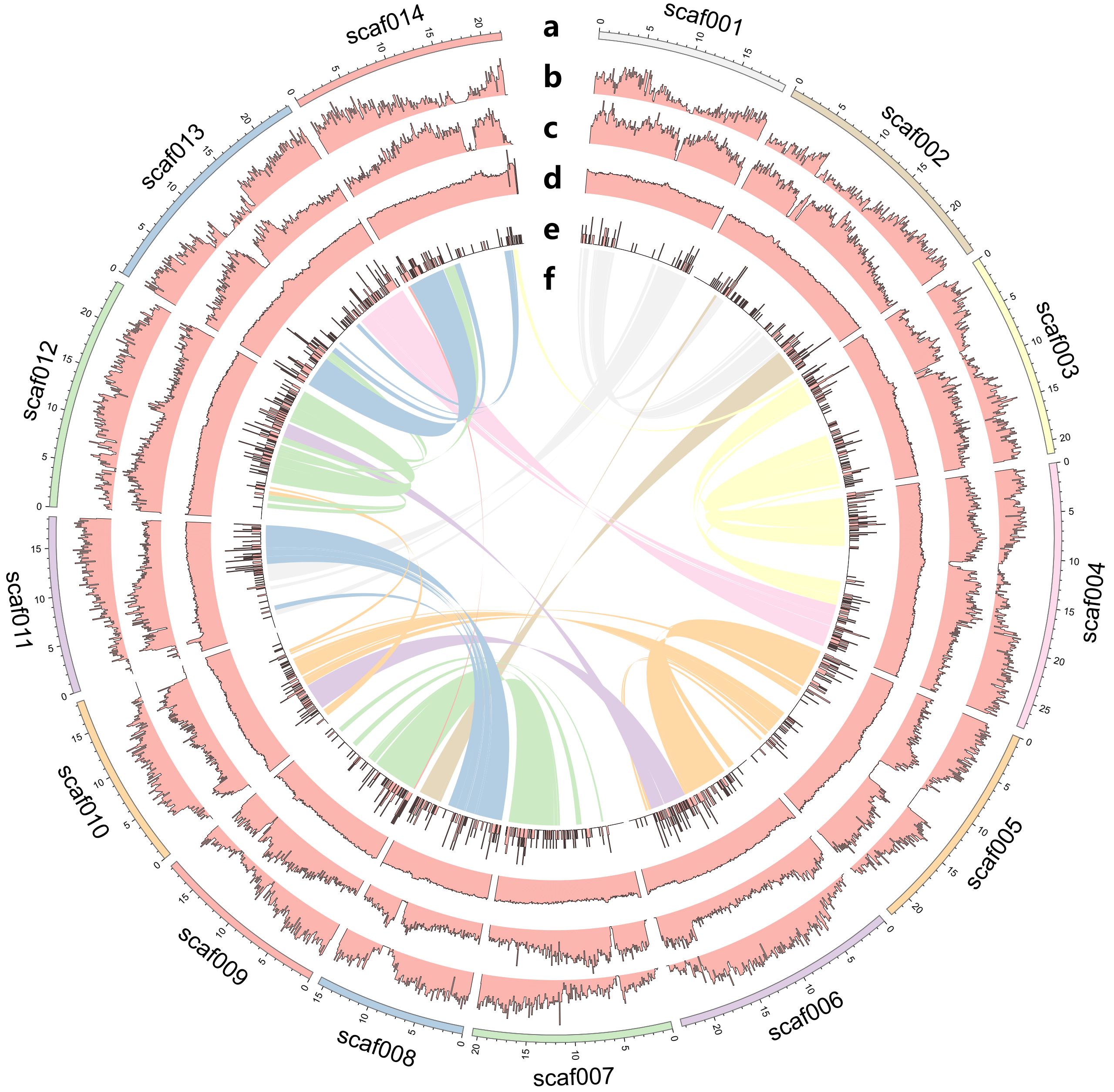

Figure 2. Intra-genomic synteny analysis and other genomic features of *Bauhinia variegata*. Tracks from outside to inside show 14 pseudo-chromosomes (a), gene density (b), transposable elements (TE) density (c), GC content (d), the density of genes located in syntenic regions (e) and intragenomic synteny (f).

3.3. Phylogenetic analyses and gene family evolution

We constructed a maximum likelihood tree for 10 legumes (*G. max*, *M. truncatula*, *L. japonicus* and *X. zambesiaca* from Papilionoideae; *A. pycnantha*, *C. fasciculata* and *G. triacanthos* from Caesalpinioideae; *B. variegata* and *C. canadensis* from Cercidoideae and *C. officinalis* from Detarioideae) based on 129 single-copy genes, with *Q. saponaria* as an outgroup. The tree topology is consistent with previous studies[1,13] and confirms that *Bauhinia* is close to *Cercis* (Fig. 3A). Interestingly, *B. variegata* has a much longer (> 3-fold) branch length than *C. canadensis* after their divergence.

Protein sequences of the 11 species were clustered into 54,370 orthogroups, with 25,927 orthogroups with two or more members. As shown in the Venn diagram (Fig. 3B), a total of 9,119 orthogroups were shared among five legumes (*B. variegata*, *C. canadensis*, *C. officinalis*, *C. fasciculata* and *G. max*), and *B. variegata* contains 732 unique orthogroups. The estimated divergence time between *B. variegata* and *C. canadensis* was 35.9 million years ago (Ma). Gene family expansion and contraction analysis identified 369 significantly expanded and 82 significantly contracted ($P < 0.05$) gene families among 4,523 expanded and 345 contracted gene families of *B. variegata*, respectively (Fig. 3A). Compared with other legumes, *B. variegata* has the second highest number of expanded genes, only lower than *G. max*. KEGG pathway enrichment analysis indicated that significantly expanded gene families were enriched in pathways of stilbenoid, diarylheptanoid and gingerol biosynthesis, flavonoid biosynthesis, cyanoamino acid metabolism, monoterpenoid biosynthesis, AGE-RAGE

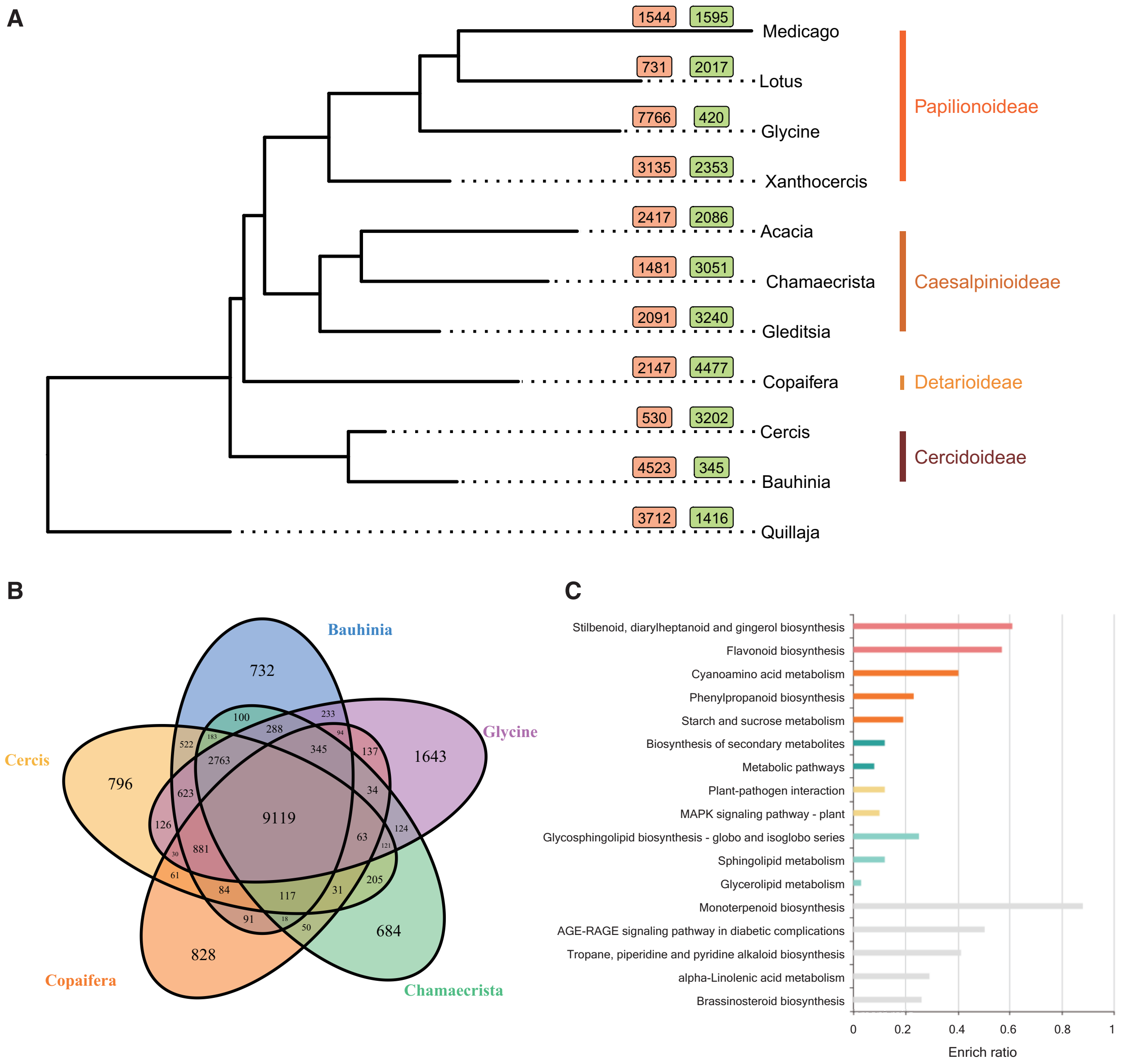

Figure 3. Phylogenomic analysis and gene family analysis. (A) Phylogenetic tree of 10 legumes and an outgroup based on concatenated sequences of 129 single-copy genes, with the numbers of expanded (left) and contracted (right) gene families shown on each branch. (B) Venn diagram showing the shared and unique gene families among five legumes. (C) KEGG pathway enrichment analysis for significantly expanded gene families in *Bauhinia variegata*. Each row represents an enriched pathway, and the length of the bar represents the enrichment ratio, which is calculated as 'input gene number'/'background gene number'. Different clusters are shown in different colours for the bar.

signalling pathway in diabetic complications, tropane, piperidine and pyridine alkaloid biosynthesis, etc. (Fig. 3C), which may contribute to its biotic and abiotic resistance, and various petal colours.

3.4. Testing the allotetraploidy origin hypothesis of *Bauhinia*

Compared with *Cercis*, which has a chromosome number of $2n = 14$, *B. variegata* has a chromosome number of $2n = 28$ (CCDB; http://ccdb.tau.ac.il/). It implies that *B. variegata* should have undergone a WGD after divergence from *Cercis*. To verify this, we searched intra-genomic syntenic blocks in the *B. variegata* genome and identified 479 intra-genomic syntenic blocks that contain 15,791 genes pairs, with the longest block containing 969 gene pairs. On average, each syntenic block contains 33 homeologous gene pairs. Collectively, these 479 syntenic blocks include 21,371 genes, indicating that 56.3% of the predicted genes of *B. variegata* exhibit synteny-based signals.

The Ks (the number of substitutions per synonymous site) distribution between gene pairs in the syntenic blocks suggests two WGDs: a young WGD at Ks = 0.22 and an old duplication at Ks = 1.74 (Fig. 4A), with the latter consistent with the γ triplication event shared in core eudicots.[52] Ks for gene pairs on syntenic blocks between *B. variegata* and *C. canadensis* exhibit two peaks of 0.14 and 0.16 (Fig. 4B), much lower than Ks (0.22) between homeologous gene pairs produced by the young WGD, suggesting the WGD might

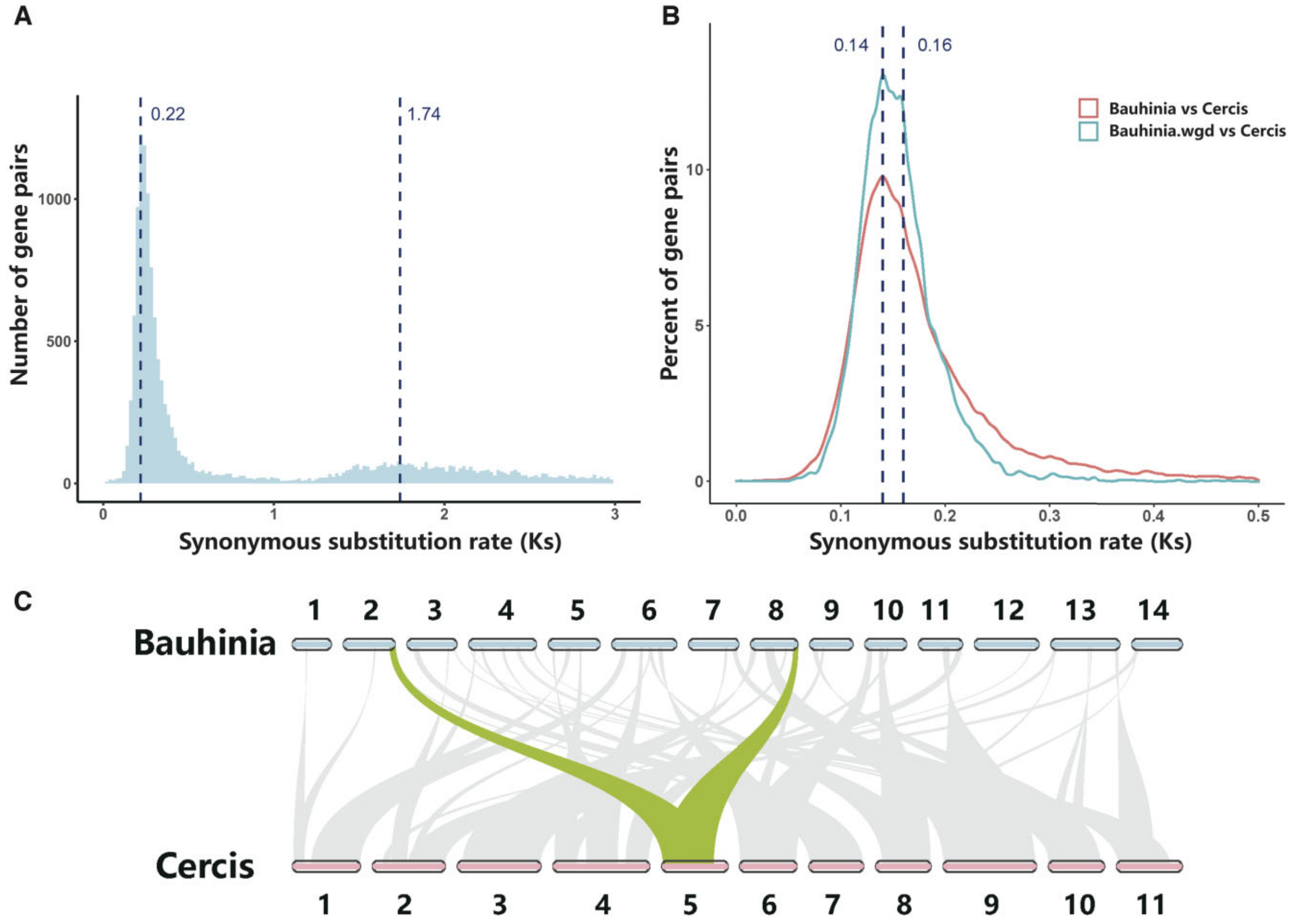

Figure 4. Identification of whole genome duplication (WGD) in *Bauhinia variegata*. (A) The histogram of synonymous substitution rate (Ks) between gene pairs on syntenic blocks in the genome of *B. variegata*. (B) The frequency density distribution of synonymous substitution rate (Ks) between *B. variegata* and *Cercis canadensis*. Shown are Ks distribution of gene pairs on syntenic blocks between the two species, and that between each of the WGD-generated duplicated genes in *B. variegata* and its corresponding ortholog in *C. canadensis*. (C) Synteny analysis between *B. variegata* and *C. canadensis*. Only 11 longest contigs of *C. canadensis* are shown here.

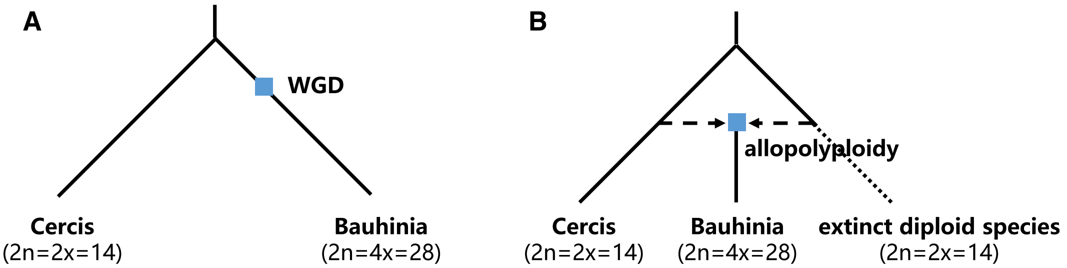

Figure 5. Alternative models for the origin of *Bauhinia*. (A) Autopolyploidy occurred in the ancestor of *Bauhinia* after divergence from *Cercis*. (B) Hybridization between the ancestor of *Cercis* and an extinct, diverged diploid species and genome doubling produced the allopolyploid ancestor of *Bauhinia*.

have occurred before the divergence between *Bauhinia* and *Cercis* if the evolutionary rates for both genera are the same. However, most syntenic regions between *B. variegata* and *C. canadensis* correspond to a rate of 2:1 (Fig. 4C), suggesting that this WGD was specific to *B. variegata*. Therefore, a greater Ks value between gene pairs produced by the young WGD might be due to accelerated evolutionary rate of *Bauhinia* after it diverged from *Cercis*, as is also shown by much longer branch length than *Cercis* on the phylogenetic tree (Fig. 3A). There are two plausible scenarios (Fig. 5) for this and both scenarios involve accelerated evolutionary rate in *Bauhinia*: one is autopolyploidy in the ancestor of *Bauhinia* and the other is allopolyploidy between a progenitor of *Cercis* and another diverged diploid species (already extinct). The latter scenario has been proposed before.[13,14] Our analyses support the latter scenario, as reasoned below.

First, the Ks distribution between each gene pairs of *B. variegata* produced by the young WGD and their corresponding ortholog of *C. canadensis* revealed two peaks at Ks = 0.14 and Ks = 0.16 (Fig. 4B), which suggests that the homeolog pairs might not originate from the same *Bauhinia* lineage. The two peaks are also consistent with those obtained from gene pairs on syntenic blocks between *B. variegata* and *C. canadensis*, suggesting these genes of this type in *B. variegata* (showing a 1:1 ratio with *Cercis*) are remnants of duplicated genes due to homeolog loss following the WGD. Second, phylogenetic analysis of 3,032 genes showed that one homeolog of *Bauhinia* was sister to the ortholog of *Cercis* rather than the other homeolog of *Bauhinia* for the majority of genes (73.9%, 75.7% and 74.7% genes when the bootstrap support values > 60, > 70 and >80 are required, respectively). This is inconsistent with the model of autopolyploidy in the ancestor of *Bauhinia*, in which the two homeologs of *Bauhinia* are expected to form sister to each other. Therefore, our genomic data support the allopolyploidy hypothesis proposed before.[13,14] Surprisingly, *C. canadensis* has a larger genome size (367 Mb[14]) than *B. variegata*, although it lacks the young WGD. We propose that genome downsizing due to genetic diploidization following the WGD in *B. variegata* can accounts for this.

4. Conclusions

We provide the first high-quality chromosome-level genome for the subfamily Cercidoideae (Leguminosae). Based on the genome sequence, we identified two WGDs in *B. variegata*, a young WGD specific to *B. variegata* and an old one corresponding to the γ triplication shared in core eudicots. Interestingly, this young WGD is not shared with *Cercis* although Ks analysis suggests so. The reason for this conflict should be accelerated evolutionary rate in *Bauhinia* after it diverged from *Cercis*, which is also supported by the much longer branch length in *B. variegata* than *C. canadensis* after their divergence. The divergence and phylogenetic analyses for each gene pairs of *B. variegata* produced by the young WGD and their corresponding ortholog in *C. canadensis* support the allopolyploidy origin hypothesis for *Bauhinia*. Consistent with the WGD, *B. variegata* possesses a large number of expanded gene families among legumes. The genome of *B. variegata* provides a valuable genomic resource for dissecting genetic basis of its ornamental traits and addressing other evolutionary and genetic questions in Cercidoideae and legumes in general.

Funding

This work was financially supported by the Natural Science Foundation of Guangdong (2021A1515010997) and Forestry Science and Technology Innovation Project of Guangdong (2018KJCX043).

Conflict of interest

None declared.

Data availability

The high-quality genome assembly and annotation of *Bauhinia variegata* have been deposited in NCBI under the accession number: JAKRYI000000000 (BioProject accession: PRJNA801801). The repeats, gene annotation and the orthogroups among 11 species obtained from OrthoFinder2 are available at https://doi.org/10.6084/m9.figshare.19298582.v1.

References

1. LPWG. 2017, A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny—the Legume Phylogeny Working Group (LPWG), *Taxon.*, **66**, 44–77.
2. Lewis, G., Schrire, B., Mackinder, B., Rico, L. and Clark, R. 2013, A 2013 linear sequence of legume genera set in a phylogenetic context—a tool for collections management and taxon sampling, *S. Afr. J. Bot.*, **89**, 76–84.
3. Yahara, T., Javadi, F., Onoda, Y., et al. 2013, Global legume diversity assessment: concepts, key indicators, and strategies, *Taxon*, **62**, 249–66.
4. Schmutz, J., Cannon, S.B., Schlueter, J., et al. 2010, Genome sequence of the palaeopolyploid soybean, *Nature*, **463**, 178–83.
5. Varshney, R.K., Chen, W., Li, Y., et al. 2012, Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers, *Nat. Biotechnol.*, **30**, 83–9.
6. Bertioli, D.J., Cannon, S.B., Froenicke, L., et al. 2016, The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut, *Nat. Genet.*, **48**, 438–46.
7. Kreplak, J., Madoui, M.-A., Cápal, P., et al. 2019, A reference genome for pea provides insight into legume genome evolution, *Nat. Genet.*, **51**, 1411–22.
8. Kamal, N., Mun, T., Reid, D., et al. 2020, Insights into the evolution of symbiosis gene copy number and distribution from a chromosome-scale *Lotus japonicus* Gifu genome sequence, *DNA Res.*, **27**, dsaa015.
9. Cui, J., Lu, Z., Wang, T., et al. 2021, The genome of *Medicago polymorpha* provides insights into its edibility and nutritional value as a vegetable and forage legume, *Hortic. Res.*, **8**, 47.
10. Griesmann, M., Chang, Y., Liu, X., et al. 2018, Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis, *Science*, **361**, eaat1743.
11. Soltis, P.S., Marchant, D.B., Van de Peer, Y. and Soltis, D.E. 2015, Polyploidy and genome evolution in plants, *Curr. Opin. Genet. Dev.*, **35**, 119–25.
12. Soltis, P.S. and Soltis, D.E. 2021, Plant genomes: markers of evolutionary history and drivers of evolutionary change, *Plants. People. Planet.*, **3**, 74–82.
13. Cannon, S.B., Mckain, M.R., Harkess, A., et al. 2015, Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes, *Mol. Biol. Evol.*, **32**, 193–210.
14. Stai, J.S., Yadav, A., Sinou, C., et al. 2019, Cercis: a non-polyploid genomic relic within the generally polyploid legume family, *Front. Plant Sci.*, **10**, 345.
15. Chen, S., Zhou, Y., Chen, Y. and Gu, J. 2018, fastp: an ultra-fast all-in-one FASTQ preprocessor, *Bioinformatics*, **34**, i884–90.
16. Xu, H., Luo, X., Qian, J., et al. 2012, FastUniq: a fast de novo duplicates removal tool for paired short reads, *PLoS One.*, **7**, e52249.
17. Marçais, G. and Kingsford, C. 2011, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers, *Bioinformatics*, **27**, 764–70.
18. 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**, 722–36.
19. Koren, S., Rhie, A., Walenz, B.P., et al. 2018, De novo assembly of haplotype-resolved genomes with trio binning, *Nat. Biotechnol.*, **36**, 1174–82.
20. Hu, J., Fan, J., Sun, Z. and Liu, S. 2020, NextPolish: a fast and efficient genome polishing tool for long-read assembly, *Bioinformatics*, **36**, 2253–5.
21. Durand, N.C., Shamim, M.S., Machol, I., et al. 2016, Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments, *Cell Syst.*, **3**, 95–8.
22. Dudchenko, O., Batra, S.S., Omer, A.D., et al. 2017, De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds, *Science*, **356**, 92–5.

23. Robinson, J.T., Turner, D., Durand, N.C., Thorvaldsdóttir, H., Mesirov, J.P. and Aiden, E.L. 2018, Juicebox. js provides a cloud-based visualization system for Hi-C data, *Cell Syst*., **6**, 256–8.e1.
24. 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**, 3210–2.
25. Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. 2005, Repbase update, a database of eukaryotic repetitive elements, *Cytogenet. Genome Res*., **110**, 462–7.
26. Flynn, J.M., Hubley, R., Goubert, C., et al. 2020, RepeatModeler2 for automated genomic discovery of transposable element families, *Proc. Natl. Acad. Sci. USA*, **117**, 9451–7.
27. Kim, D., Paggi, J.M., Park, C., Bennett, C. and Salzberg, S.L. 2019, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype, *Nat. Biotechnol*., **37**, 907–15.
28. Li, H., Handsaker, B., Wysoker, A., et al.; 1000 Genome Project Data Processing Subgroup. 2009, The sequence alignment/map format and SAMtools, *Bioinformatics*, **25**, 2078–9.
29. Kovaka, S., Zimin, A.V., Pertea, G.M., Razaghi, R., Salzberg, S.L. and Pertea, M. 2019, Transcriptome assembly from long-read RNA-seq alignments with StringTie2, *Genome Biol*., **20**, 278.
30. Ou, S., Su, W., Liao, Y., et al. 2019, Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline, *Genome Biol*., **20**, 275.
31. Stanke, M. and Morgenstern, B. 2005, AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints, *Nucleic Acids Res*., **33**, W465–7.
32. Jones, P., Binns, D., Chang, H.-Y., et al. 2014, InterProScan 5: genome-scale protein function classification, *Bioinformatics*, **30**, 1236–40.
33. Törönen, P., Medlar, A. and Holm, L. 2018, PANNZER2: a rapid functional annotation web server, *Nucleic Acids Res*., **46**, W84–8.
34. Schwacke, R., Ponce-Soto, G.Y., Krause, K., et al. 2019, MapMan4: a refined protein classification and annotation framework applicable to multi-omics data analysis, *Mol. Plant*., **12**, 879–92.
35. Quinlan, A.R. and Hall, I.M. 2010, BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics*, **26**, 841–2.
36. Krzywinski, M., Schein, J., Birol, I., et al. 2009, Circos: an information aesthetic for comparative genomics, *Genome Res*., **19**, 1639–45.
37. Emms, D.M. and Kelly, S. 2019, OrthoFinder: phylogenetic orthology inference for comparative genomics, *Genome Biol*., **20**, 238.
38. Löytynoja, A. and Goldman, N. 2008, Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis, *Science*, **320**, 1632–5.
39. Suyama, M., Torrents, D. and Bork, P. 2006, PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments, *Nucleic Acids Res*., **34**, W609–12.
40. Darriba, D., Posada, D., Kozlov, A.M., Stamatakis, A., Morel, B. and Flouri, T. 2020, ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models, *Mol. Biol. Evol*., **37**, 291–4.
41. Kozlov, A.M., Darriba, D., Flouri, T., Morel, B. and Stamatakis, A. 2019, RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference, *Bioinformatics*, **35**, 4453–5.
42. Yang, Z. 2007, PAML 4: phylogenetic analysis by maximum likelihood, *Mol. Biol. Evol*., **24**, 1586–91.
43. Mendes, F.K., Vanderpool, D., Fulton, B. and Hahn, M.W. 2021, CAFE 5 models variation in evolutionary rates among gene families, *Bioinformatics*, **36**, 5516–8.
44. Bu, D., Luo, H., Huo, P., et al. 2021, KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis, *Nucleic Acids Res*., **49**, W317–25.
45. Altschul, S.F., Madden, T.L., Schäffer, A.A., et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*., **25**, 3389–402.
46. Wang, Y., Tang, H., DeBarry, J.D., et al. 2012, MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity, *Nucleic Acids Res*., **40**, e49.
47. Wang, D., Zhang, Y., Zhang, Z., Zhu, J. and Yu, J. 2010, KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies, *Genomics Proteomics Bioinformatics*, **8**, 77–80.
48. Wickham, H. 2016, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag: New York.
49. Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M. and Paterson, A.H. 2008, Synteny and collinearity in plant genomes, *Science*, **320**, 486–8.
50. Katoh, K. and Standley, D.M. 2013, MAFFT multiple sequence alignment software version 7: improvements in performance and usability, *Mol. Biol. Evol*., **30**, 772–80.
51. Zhang, Z., Xiao, J., Wu, J., et al. 2012, ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments, *Biochem. Biophys. Res. Commun*., **419**, 779–81.
52. Jiao, Y., Leebens-Mack, J., Ayyampalayam, S., et al. 2012, A genome triplication associated with early diversification of the core eudicots, *Genome Biol*., **13**, R3.