Journal Pre-proof



Rosaceae phylogenomic studies provide insights into the evolution of new genes

Lan Jiang, Xiaoxu Li, Kun Lyu, Han Wang, Zhiyuan Li, Wang Qi, Lin Zhang, Yunpeng Cao

PII: S2468-0141(24)00069-4

DOI: https://doi.org/10.1016/j.hpj.2024.02.002

Reference: HPJ 590

To appear in: *Horticultural Plant Journal*

Received Date: 17 December 2023

Revised Date: 29 January 2024

Accepted Date: 27 February 2024

Please cite this article as: Jiang L, Li X, Lyu K, Wang H, Li Z, Qi W, Zhang L, Cao Y, Rosaceae phylogenomic studies provide insights into the evolution of new genes, *Horticultural Plant Journal*, https://doi.org/10.1016/j.hpj.2024.02.002.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Copyright © 2024 Chinese Society for Horticultural Science (CSHS) and Institute of Vegetables and Flowers (IVF), Chinese Academy of Agricultural Sciences (CAAS). Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd.

Journal Pre-proof

Rosaceae phylogenomic studies provide insights into the evolution of new genes

Lan Jiang[a,#], Xiaoxu Li[b,#], Kun Lv[a,#], Han Wang[f], Zhiyuan Li[e], Wang Qi[e], Lin Zhang[g], and Yunpeng Cao[c,d,*]

Running title: The evolution of new genes

[a]*Anhui Province Key Laboratory of Non-coding RNA Basic and Clinical Transformation (Wannan Medical College), Central Laboratory, Yijishan Hospital of Wannan Medical College, Wuhu 241000, China.*

[b]*Beijing Life Science Academy, Beijing 102209, China.*

[c]*CAS Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan 430074, China*

[d]*Center of Economic Botany, Core Botanical Gardens, Chinese Academy of Sciences, Wuhan 430074, China*

[e]*Key Laboratory for Tobacco Gene Resources, Tobacco Research Institute, Chinese Academy of Agricultural Sciences, Qingdao, China*

[f]*School of Life Sciences, Anhui Agricultural University, Hefei 230036, China*

[g]*School of Basic Medical Sciences, Hubei University of Chinese Medicine, Wuhan 430065, China*

[#]these authors contributed equally to this work

*Corresponding authors: Yunpeng Cao (email: caoyunpeng@wbgcas.cn and xfcypeng@126.com)

Orcid: Yunpeng Cao, 0000-0001-5976-2382

Orcid: Lan Jiang, 0000-0003-2778-1081

CRediT authorship contribution statement

LJ and YC designed the study. LJ, YC, XL, HW, KL, and LZ performed the analyses of the genome, and transcriptome sequence. WQ, LZ and YC collected the samples. LJ, XL and YC performed the experiments. YC wrote the manuscript. LJ and YC revised the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We thank Wuhan Onemore-tech Co., Ltd. for their assistance with genome sequencing and analysis, and also extend our thanks to the editors and reviewers for their careful reading and helpful comments on this manuscript. This work was supported by the National Natural Science Foundation of China (Grant No. 32201602, 82304680), the Natural Science Fund of Hubei Province (Grant No. 2023AFB1036), and the Program for Excellent Sci-tech Innovation Teams of Universities in Anhui Province (Grant No: 2022AH010074), and Anhui Provincial Natural Science Foundation (2308085QH295), and Natural Science Research Project of Anhui Educational Committee (Grant Number: 2023AH040259), and the Talent Scientific Research Start-up Foundation, Wannan Medical College (Grant Number: YR20230110).



Transystom
Editing & Translation

Editorial Certificate

This is to certify that the manuscript detailed below was edited for proper usage of scientific English by one or more native English-speaking editors at Transystom Editing and Translation Services. A copy of the edited manuscript is available for the journal or publisher upon request. To verify this certificate, please contact us at info@transystom.com and provide the verification code provided below.

Manuscript Title: Rosaceae phylogenomic studies provide insights into the evolution of new genes

Authors: Lan Jiang, Xiaoxu Li, Kun Lv, Han Wang, Zhiyuan Li, Wang Qi, Lin Zhang, and Yunpeng Cao

Date Issued: January 27, 2024

Verification Code: 829.20240118.SE

The authors may choose to accept or reject suggestions and changes to the document after our editing process is completed and prior to submission to any journal. Thus, this certificate allows the editor to identify changes introduced to the manuscript after editing. If you have any questions or concerns, please feel free to contact us.

Transystom provides editing service for Chinese researchers who wish to publish their manuscripts in international journals. The editing service is provided by experienced native English-speaking editors with expertise in specific areas of research, as well as translation staff who are native Chinese speakers with good command of English and who will help to ensure that the intended meaning of poorly written text is retained.

Journal Pre-proof

Rosaceae phylogenomic studies provide insights into the evolution of new genes

Lan Jiang[a,1], Xiaoxu Li[b,1], Kun Lyu [a,1], Han Wang[f], Zhiyuan Li[e], Wang Qi[e], Lin Zhang[g], and Yunpeng Cao[c,d,*]

[a]*Anhui Province Key Laboratory of Non-coding RNA Basic and Clinical Transformation (Wannan Medical College), Central Laboratory, Yijishan Hospital of Wannan Medical College, Wuhu, Anhui 241000, China*

[b]*Beijing Life Science Academy, Beijing 102209, China*

[c]*CAS Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, Hubei 430074, China*

[d]*Center of Economic Botany, Core Botanical Gardens, Chinese Academy of Sciences, Wuhan, Hubei 430074, China*

[e]*Key Laboratory for Tobacco Gene Resources, Tobacco Research Institute, Chinese Academy of Agricultural Sciences, Qingdao, Shangdong 266000, China*

[f]*School of Life Sciences, Anhui Agricultural University, Hefei, Anhui 230036, China*

[g]*School of Basic Medical Sciences, Hubei University of Chinese Medicine, Wuhan, Hubei 430065, China*

[1] These authors contributed equally to this work.
*Corresponding authors.
Email address: caoyunpeng@wbgcas.cn; xfcypeng@126.com

Journal Pre-proof

Abstract

Rosa banksiae, known as Lady Banks' rose, is a perennial ornamental crop and a versatile herb in traditional Chinese medicine. Given the lack of genomic resources, we assembled a HiFi and Nanopore sequencing-derived 458.58 Mb gap-free telomere-to-telomere high-quality *R. banksiae* genome with a scaffold N50 = 63.90 Mb. The genome of *R. banksiae* exhibited no lineage-specific whole-genome duplication compared with other Rosaceae. The phylogenomic analysis of 13 Rosaceae and *Arabidopsis* through a comparative genomics study showed that numerous gene families were lineage-specific both before and after the diversification of Rosaceae. Some of these genes are candidates for new genes that have evolved from parental genes through fusion events. Fusion genes are divided into three types: Type-I and Type-II genes contain two parental genes that are generated by duplication, distributed in the same and different regions of the genome, respectively; and Type-III can only be detected in one parental gene. Here, Type-I genes are found to have more relaxed selection pressure and lower Ks values than Type-II, indicating that these newly evolved Type-I genes may play important roles in driving phenotypic evolution. Functional analysis exhibited that newly formed fusion genes can regulate the phenotype traits of plant growth and development, suggesting the functional significance of these genes. This study identifies new fusion genes that could be responsible for phenotype evolution and provides information on the evolutionary history of recently diverged species in the *Rosa* genus. Our data represents the major progress in understanding the new fusion genes evolution pattern of Rosaceae and provides an invaluable resource for phylogenomic studies in plants.

***Keywords*:** *Rosa banksiae*; Gap-free genome; Rosaceae; Phenotype; Evolutionary patterns

Journal Pre-proof

1. Introduction

The genus *Rosa* belongs to the Rosaceae family (Van Huylenbroeck, 2018). The family of Rosaceae is divided into four subfamilies: Maloideae, Amygdaloideae, Spiraeoideae, and Rosoideae which contains the genus *Rosa* (Longhi et al., 2014). Plants of the genus *Rosa* are usually shrubs with sticky stigmas, abundant styles, numerous stamens, five petals, and five sepals, and are described as semi-woody perennials (Leus et al., 2018). *Rosa banksiae* f. *lutea* (Lindl.) Rehd, a species of the genus *Rosa* family, is one of the most important deciduous shrub flowering plants, native to central and western China (Zhuang et al., 2021). It is also a medicinally, ecologically, and economically important plant with many characteristics such as cold endurance, drought resistance, and strong resistance to barren land(Zhuang et al., 2021).

The availability of genome-wide information provides new opportunities for the study of genomics and genetics to understand the genetic regulation and molecular mechanisms of different phenotypes (Shen et al., 2021; Deng et al., 2022; Jiang et al., 2022). High-quality, complete, accurate, and gap-free reference genomes are essential for gene discovery and traits (Natarajan et al., 2021; Aganezov et al., 2022; Deng et al., 2022). Zhang et al. (2019a) identified a specific LTR retrotransposon associated with apple red skin in the homozygous line HFTH1 genome, but not in the genome of GDDH13, using high-quality genome assembly. At this point, researchers have sequenced and published more than one *de novo* genome in humans and many plants, such as rice, *Chimonanthus praecox*, maize, *Arabidopsis*, *Bletilla striata* (Song et al., 2021; Navratilova et al., 2022; Nurk et al., 2022; van Rengs et al., 2022; Wang et al., 2022; Zhang et al., 2022). The large amount of information provided by the sequencing of multiple genomes is crucial for further research on countless important agronomic traits.

The family of Rosaceae contains many important horticultural or fruit tree plants, and many genome structures have been studied, such as pear, peach, Chinese plum, strawberry, apple, and rose (Shulaev et al., 2011; Wu et al., 2013; Raymond et al., 2018; Zhang et al., 2019a; Zheng et al., 2022). However, each genome still has multiple gaps and is incomplete. Mutant collections generated from the same genetic pool, combined with high-quality reference genomes, will facilitate the isolation and discovery of desirable mutants required for breeding and genetics. The ultimate goal of genome assembly is the gap-free genome, which opens up new opportunities for the identification of structural variants (SVs) and unique genes in "dark matter" regions, such as segmental duplications, transposable elements (TEs), centromeres (Li et al., 2021; Natarajan et al., 2021; Song et al., 2021; Zhang et al., 2022). Long-read technologies are the standard for producing high-quality assemblies, especially for complex genomes such as horticultural crop genomes. For example, the long fragment reads were provided by PacBio HiFi reads with 99% accuracy, which facilitates accurate assembly of highly repetitive regions, such as telomeres and centromeres. While the impact of these technologies on genome assembly is enormous, their maturity for constructing complete chromosomal sequences from telomere to telomere remains to be further addressed (Michael et al., 2018). In recent, the Telomere-to-Telomere (T2T) consortium assembled the first T2T assembly of the X chromosome in human (Miga et al., 2020). This assembly combines multiple sequencing technologies: including single-molecule real-time (SMRT) sequencing, Oxford Nanopore Technology (ONT) sequencing, Bionano Genomics (BNG), and 10X Genomics (10X). Although the final assembly obtained by these techniques is continuous, there are still some gaps, and the complete X chromosome is finally obtained by artificial hole filling. This gigantic effort is impossible for most genome projects because it is too time-consuming and expensive. However, the development of sequencing technology has provided the possibility to solve the assembly of T2T (Wenger et al., 2019). Gap-free genomes have been sequenced in banana, watermelon, human, rice, and *Arabidopsis* by combing chromosome conformation capture (Hi-C), ONT, and high-fidelity (HiFi) sequencing technologies (Song et al., 2021; Navratilova et al., 2022; Nurk et al., 2022; van Rengs et al., 2022;

Wang et al., 2022; Zhang et al., 2022), but research on gap-free genomes in Rosaceae is still excluded.

New genes provide genomic novelty for the evolution of organisms, which can be evolved from several mechanisms, such as gene duplication, gene fission, gene fusion, and *de novo* evolution (Chen et al., 2013; Long et al., 2013; Song et al., 2022). The development of whole-genome sequencing has greatly facilitated the investigation of new gene occurrences, especially those arising from parental gene fusions, as related homologs can be traced within a specific clade. Recently, the retrotransposed gene *Adh*, which encodes alcohol dehydrogenase, was inserted into the duplicated gene *Ynd* to create the well-known new gene *jingwei* in *Drosophila* (Long et al., 2003). In plants, fusion genes play a crucial role in influencing phenotypes and generating novel functions. For example, the new gene *EXOV* in *Arabidopsis*, resulting from a partial gene duplication of its parental gene *EXOVL*, has shown rapid evolution of new functions and novel phenotypic effects (Huang et al., 2022); *CHANNEL11/12* in *Arabidopsis* has been identified to regulate multiple pathogen-resistance responses(Yoshioka et al., 2006); the fusion genes *Osjap07g28390* and *Osjap09g15430* in *Oryza sativa* L. v.g. *japonica* were associated with drought resistance and had significant phenotypic effects (Zhou et al., 2022); *ATP6* and *Rfo* have been found to induce male sterility in *O. sativa* and *Raphanus raphanistrum*, respectively (Brown et al., 2003; Wang et al., 2006); *BS1* in *Zea mays* is involved in the normal reproductive development (Elrouby and Bureau, 2010). Here, we sequenced the first reference genome of *R. banksiae* and obtained a T2T gap-free assembly level. Through the high-quality assembly of the *R. banksiae* genome, we found hundreds of lost and gained gene families during the evolution of Rosaceae, shedding light on previously unidentified new genes. The well-annotated genes from *R. banksiae* serve as a reliable reference for numerous phylogenetic and phylogenomic research, and traditional Chinese medicine.

2. Materials and methods

2.1. Plant materials and genomic DNA sample preparation

Fresh young leaves of *R. banksiae* used for T2T genome sequencing were obtained from Wannan Medical College (Wuhu, Chinese). The Qiagen Genomic DNA Kit (Qiagen, Valencia, CA) was used to extract the high-molecular weight DNA following the manufacturer's manual. The quantity and quality of extracted DNA were assessed using the Qubit 3.0 Fluorometer (Invitrogen Life Technologies, Carlsbad, CA) and the NanoDrop One UV–Vis spectrophotometer (ThermoFisher Scientific, Waltham, MA), respectively. Finally, the large DNA fragments were recovered by gel cutting using the Blue Pippin system (Sage Science, Beverly, MA).

2.2. ONT library preparation and sequencing

For the ultra-long Nanopore library, we selected about 10 μg of genomic DNA (> 100 kb) using the SageHLS HMW library system (Sage Science) and then constructed the sequencing libraries with the Ligation sequencing 1D Kit (Oxford Nanopore Technologies, Oxford, UK), based on the standard protocol of ONT. Using the PromethION sequencer (ONT, UK), we sequenced DNA libraries (about 800 ng) at the Genome Center of Grandomics (Wuhan, China). Finally, the ONT Guppy software (v1.8) was used to perform base calling on fast5 files, and the high-quality reads (i.e. "passed filter" data) were selected for further analysis.

2.3. HiFi sequencing and assembly

Based on the manufacturer's instructions of PacBio, we constructed the SMRTbell libraries with 15 kb preparation solutions (Pacific Biosciences, CA). In brief, we used the gTUBEs to shear the approximately 15 μg genomic DNA sample, removed single-strand overhangs, and then performed A-tailed, end-repaired, and damage

Journal Pre-proof

repaired. The hairpin adapters were ligated with the fragments for HiFi sequencing. The nuclease from SMRTbell Enzyme Cleanup Kit was used to treat the library. Target fragments were screened and purified using BluePippin (Sage Science), and the SMRTbell library was similarly processed with AMPure PB beads. The PacBio Sequel II instrument was used to perform the sequencing with Sequel II Binding Kit 2.0 and Sequencing Primer V2, as described by Jiang et al. (2022). The HiFiasm v0.14-r312 was used to assemble the genome using HiFi reads with default parameters (Cheng et al., 2021). Finally, we converted sequence graphs in the GFA to FASTA format using gfatools (https://github.com/lh3/gfatools).

2.4. Hi-C sequencing and scaffolding

The Hi-C library was prepared based on a published protocol (Jiang et al., 2022). Briefly, 1% formaldehyde solution was used to cross-link 2 g of young leaves *in situ*. New England Biolabs (MboI) was used to extract and digest chromatin. We enriched and quality-checked DNA fragments, and then we constructed and sequenced the sequencing libraries using Illumina NovaSeq 6000 by 150-bp paired-end mode. Finally, 58.05 Gb of Hi-C reads were produced with 124× coverage. HicPro v2.11.1 was used to filter the Hi-C sequenced data (Servant et al., 2015), and then Juicer v1.6 was used to align the clean data to the corrected contig genome (Durand et al., 2016). 3D-DNA v180922 was used to further construct a draft assembly genome using the mapped data (Dudchenko et al., 2017), and the Juicebox v1.11.08 was further used to correct the obtained draft genome (Robinson et al., 2018).

2.5. Replacing HiFi assemblies with ONT reads

TGS-GapCloser v1.2.0 was used to fill the remaining gaps between contigs with parameter: --min_nread 10, and extend the contigs using the coverage relationship between the ONT ultra-long data and the assembled contigs (Xu et al., 2020). Subsequently, the extended genome was error-corrected using next-generation sequencing data. For the centromere of *R. banksiae*, the ONT reads longer than 100 kb were first aligned to the extended genome using Winnowmap v1.11 and further obtained reads aligning to the centromeric higher-order repeat array (Jain et al., 2020). Finally, TandemQUAST was used to detect the centromeric sequences with the following procedure: tandemquast.py - o {out_dir} --nano {ont_reads.fa} -t 36 genome.fa (Mikheenko et al., 2019).

2.6. Genome assembly quality assessment

We evaluated the genome quality using various methods. A lower homozygosity rate indicates higher genome accuracy, while a higher heterozygosity rate suggests increased genetic diversity in the genome. The sequencing reads obtained were aligned to the genome using BWA v0.7.8. Subsequently, SNP sites were identified using SAMtools v1.10, Picard v2.25.0, and GATK v3.8.0 software (McKenna et al., 2010). Finally, we calculated the rates of homozygosity and heterozygosity for SNP and InDel, respectively. The structure variation (SV) analysis was used to verify the completeness of repeat regions and correct assembly. Specifically, minimap2 v2.1 and Winnowmap v1.11 were used to align HiFi and ONT ultra-long reads to the assembly. The SVs were called using NextSV v3.1.0, which combines the results from sniffles2 v2.2 and cuteSV v2.0.3 (Jain et al., 2020; Jiang et al., 2020; Smolka et al., 2022). This approach can assess the presence of substantial insertions and deletions within repeat regions. The completeness and quality of the *R. banksiae* genome assembly and annotation were evaluated using BUSCO v3.0.2 with the embryophyta_odb10 database and default parameters (Simao et al., 2015).

Journal Pre-proof

2.7. Genome annotation

We annotated the *R. banksiae* genome using genomic sequences, as well as gene function information, gene structure information, repeated sequences, and non-coding RNAs. RepeatMasker and RepeatProteinMask were used to annotate the repeated sequences using the Repbase database (Chen, 2004). TRF, LTR-FINDER, RepeatModeler, and Repeatscout were carried out to detect and annotate repeated sequences *de novo* (Price et al., 2005; Xu and Wang, 2007; Flynn et al., 2020). The annotation of gene structures was used in three methods. First, GLIMMER HMM, AUGUSTUS, and GENSCAN were used for *de novo* predictions based on the *R. banksiae* genome (Burge and Karlin, 1997; Benson, 1999; Majoros et al., 2004; Stanke et al., 2006). Second, GENEWISE was used for homologous annotation by searching five related species: *Rubus ccidentalis*, *R. wichuraiana*, *R. chinensis*, *Prunus persica*, and *Arabidopsis thaliana* (Madeira et al., 2019). Next, we performed transcript annotations based on the RNA-Seq data by HISAT2, PASA, and STRINGTIE (Haas et al., 2003; Pertea et al., 2015; Kim et al., 2019). The MAKER package v2 was further used to merge all gene models (Cantarel et al., 2008). Finally, we manually filtered and corrected the genes that contained transposable element (TE) sequences in coding regions using IGV-GSAman v0.6.83 (https://gitee.com/CJchen/IGV-sRNA). Gene functions were annotated using BLASTP by leveraging protein databases such as NCBI non-redundant, Gene Ontology, InterPro, TrEMBL, COG/KOG, KEGG, and SwissProt (Boeckmann et al., 2003; Kanehisa et al., 2008; Jones et al., 2014; Gene Ontology, 2019; Zhou et al., 2023). Non-coding RNAs, including tRNA, snRNAs, miRNAs, ncRNAs, and rRNA, were annotated by the Rfam database (https://rfam.org/), tRNAScan-SE v2.0 and RNAmmer v 1.2 (Lagesen et al., 2007; Kalvari et al., 2018; Chan and Lowe, 2019).

2.8. Comparative genomics

OrthoFinder v2.5.5 was used to obtain the orthologous groups from the *R. banksiae* genome and 13 other angiosperms genomes (Emms and Kelly, 2019). The cluster granularity setting was carried out using the MCL inflation of default parameters (1.5), and alignment was performed in Diamond v2.1.8 (Buchfink et al., 2015). IQ-tree v2.2.2.6 was used to construct the species tree and gene trees of all the orthologous groups (Nguyen et al., 2015). The single-copy genes were selected to further obtain the evolutionary relationships among these genomes, after which a species tree was built with the best model the Jones-Taylor-Thornton (JTT) model in RAXML v.8.2.4 (Stamatakis, 2014). The MCMCTREE included in PAML v.4.9 was used to calculate the divergence times of plant species (Yang, 2007; Wang et al., 2023;). We selected four calibration points from the TimeTree website as normal priors to reduce age (Kumar et al., 2017), with reference to speciation times of 14-48.5 million years ago (Ma) for the divergence between *P. avium* and *P. persica*, 1.1-2.26 Ma for that between *M. domestica* and *P. communis*, 28.3-68.8 Ma for that between *R. occidentalis* and *R. multiflora*, and 102-113.8 Ma for that between *A. thaliana* and *R. chinensis*. Based on the divergence predicted by the phylogenetic tree, we identified the contraction and expansion of gene families in these genomes using CAFE v4.2 (Han et al., 2013), as described by Nowak et al. (2021). The WGDI v0.6.5 was used to search the paralogous and orthologous gene pairs and syntenic blocks (Sun et al., 2022). The NGenomeSyn v1.41 software (https://github.com/hewm2008/NGenomeSyn) was used to plot the synteny figure. WGD events were inferred and the synonymous substitutions per synonymous site (Ks) values were calculated according to the procedures described for the *Bletilla striata* genome (Jiang et al., 2022).

2.9. Positive selection tests

The branch-site models, as implemented in the codeml module of the PAML package, are widely employed

Journal Pre-proof

for identifying positively selected genes (PSGs), as documented by Nowak et al. (2021). Thus, the PSGs in the *R. banksiae* genome was identified within the single-copy orthologs among the Rosaceae species (*P. avium, P. persica, M. domestica, P. communis, P. bretschneideri, P. betulifolia, F. vesca, R. banksiae, R. rugosa, R. chinensis, R. wichuraiana*, and *R. multiflora*) using GWideCodeML v1.1 (Macías et al., 2020). Subsequently, we defined *R. banksiae* as the 'foreground' phylogeny, and set the other Rosaceae species as the 'background' phylogeny. A neutral branch site model (omega = 1, fix_omega = 1, NSsites = 2, and Model = 2) and an alternative branch-site model (fix_omega = 0, NSsites = 2, and Model = 2) were tested. Genes exhibiting Bayesian empirical Bayes (BEB) sites with a probability greater than 90% and a corrected *P*-value less than 0.1 were identified as having undergone positive selection.

2.10. RNA-Seq data analysis

Five tissues [stamen, stem, leaf, root, and Pistil and Ovary (PO)] of *R. banksiae* for RNA-Seq data analysis were obtained from Wannan Medical College (Wuhu, Chinese). Ten grams of each tissue type were collected from the seedlings (stem, leaf, root). The RNAprep Pure Plant Kit (Tiangen, Beijing, China) was used to extract total RNA. Complementary DNA (cDNA) libraries were created using high-quality total RNA in accordance with the manufacturer's instructions. Raw reads were produced by the Beijing Institute (BGI) BGISEQ-500 platform. Trimmomatic v0.39 was used to process the RNA-Seq data (Bolger et al., 2014), and HISAT2 v2.2.1 was used to map these obtained clean reads into *R. banksiae* genome (Pertea et al., 2016; Xin et al., 2023). StringTie v2.2.1 was used to estimate the normalized transcripts per million (TPM) value (Pertea et al., 2015). A gene was deemed to be expressed in our study if its expression value exceeded 0 TPM in at least one sample, as described by Zhou et al. (2022). The τ (Tau, tissue specificity index) was calculated using TPM values. If $\tau \geq 0.9$, it indicates that this gene is tissue-specific expressed (Mehta et al., 2021).

2.11. Functional analysis of Rb05g038170.1

We cloned the coding sequence (CDS) of *Rb05g038170.1* using the primers: ATGACTGAGCTCCCAATGGC and TTCCATAGAGTGTCTGCCAACACAT, and inserted them into the pCAMBIA1304 overexpression vector (Clontech, Beijing, China) by the homologous recombinant method using the GenRec Assembly Master Mix Kit (Generalbiol, Chuzhou, China). The suspension of *Agrobacterium tumefaciensstrain* GV3101 harboring pCAMBIA1304 vector with *Rb05g038170.1* were injected into the leaves of 3-week-old tobacco (*Nicotiana benthamiana*) by the *Agrobacterium*-mediated method, as described by Jiang et al. (2022). The pCAMBIA1304 was served as the control. The Confocal laser scanning microscopy (CarlZeiss LSM710, Germany) was used to detect GFP signals of the expressed pRb05g038170.1-GFP two days after infiltration. To acquire T_1 seeds, the half-strength MS media containing 20 mg · L^{-1} of hygromycin was used to screen the T_0-generation seeds. The PCR and sequencing was further used to confirm the positive transformants.

3. Results

3.1. Genome sequencing and assembly of R. banksiae

A diploid (2n= 2x= 14) *R. banksiae* f. *lutea* (Lindl.) Rehd plant was used for whole-genome sequencing and T2T gap-free chromosome-level assembly via PacBio HiFi reads, ONT ultra-long reads sequencing, and Illumina short-read sequencing. The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive (Chen et al., 2021b) in National Genomics Data Center(Members, 2023), China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences (GSA: CRA007289) that are

publicly accessible at https://ngdc.cncb.ac.cn/gsa. Using K-mer analysis (Fig. S1), the size of the *R. banksiae* genome was estimated to be 448 Mb. The extremely low heterozygosity level of this inbred line was determined to be 0.7%, proving to be well suited for genome assembly without phasing the two haplotypes. Here, the PacBio HiFi reads were used to assemble the *R. banksiae* genome. Briefly, we sequenced 26.83 Gb HiFi reads with the N50 size of 14.53 Kb and then used these reads for preliminary assembly using HiFiasm. This assembly, RbGA_v0.1, consisted of 149 contigs with an N50 of 54.86 Mb. After this step, 58.05 Gb of Hi-C chromosome conformation capture data supported all contigs being oriented, ordered, and anchored to give 7 chromosomes using 3D-DNA (i.e. RbGA_v0.2 assembly). Then, we further sequenced 22.07 Gb ONT ultra-long reads (>100 Kb) with the N50 size of 100.57 Kb and combined it with RbGA_v0.2 to improve the assembly (i.e. RbGA_v0.3). Subsequently, the gaps in the genome were filled using ONT reads, followed by the use of Illumina reads to polish the genome assembly. Finally, gap-filling and error corrections in RbGA_v0.3 generated a 458.58 Mb reference genome (RbGA_v1) with a contigs N50 = 63.90 Mb and a scaffold N50 = 63.90 Mb (Table 1). The finalized RbGA_v1 assembly had 7 scaffolds corresponding to 7 chromosomes. Taken together, we obtained a telomere-to-telomere gap-free genome with the size of 458.58 Mb and all chromosomes contained seven-base telomeric repeat unit (TTTAGGG at the 3′ end or CCCTAAA at the 5′ end) on one end (Fig. 1).

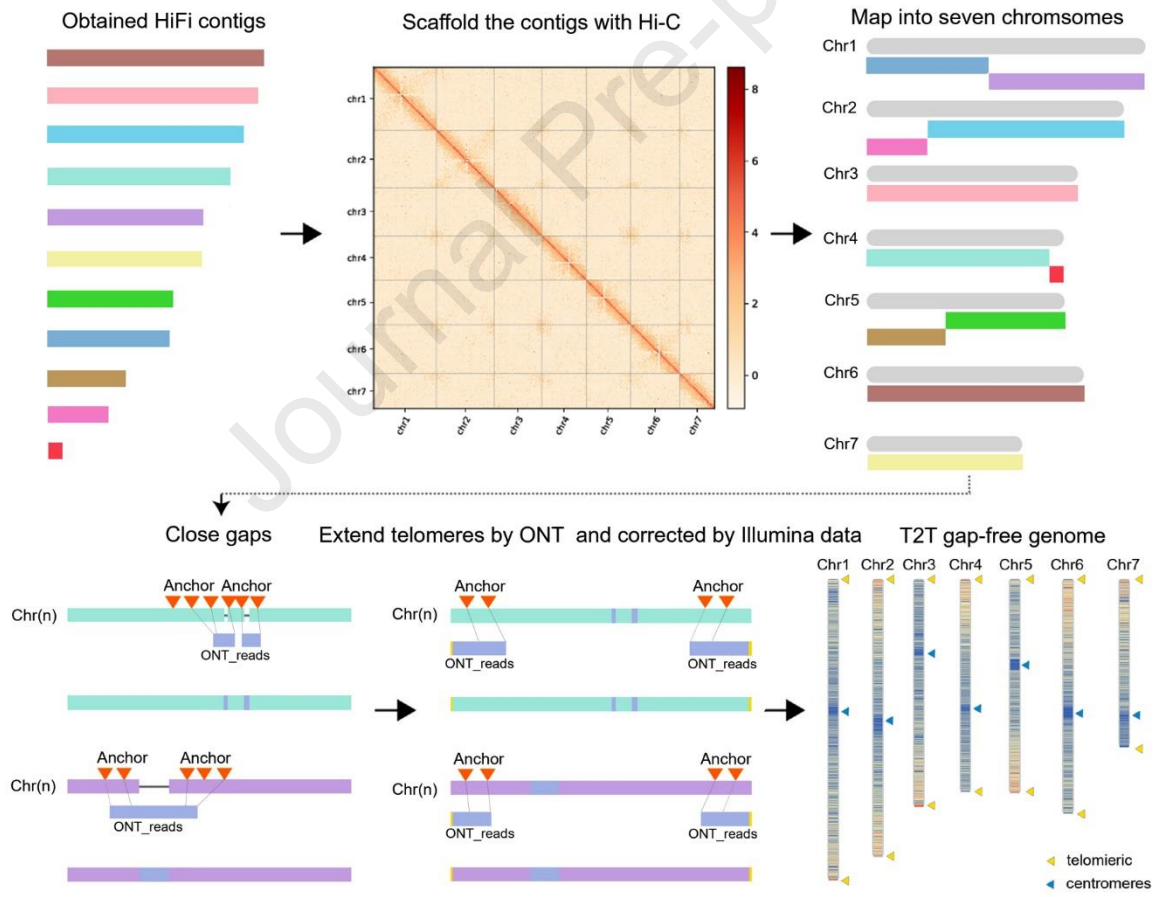


Fig. 1 High-quality T2T gap-free genome assembly of *R. banksiae*

Eleven contigs were obtained by assembly of HiFi reads. Using Hi-C data, the 11 contigs were oriented and ordered to seven scaffolds. Gaps were filled and telomeres were extended using ONT ultra-long reads based on BAC anchors. The yellow arrows represented telomeres.

Journal Pre-proof

Table 1 Summary of the *R. banksiae* genome assembly and annotation

Assembly	
Number of chromosomes	7
Assembly length (Mb)	458.58
Contig N50 (Mb)	63.90
Scaffold N50 (Mb)	63.90
BUSCOs (%)	98.33
Length of Ns	0
Rate of Ns	0
GC content of the genome (%)	38.78
Annotation	
Number of predicted protein-coding genes	31 629
Median gene length (bp)	2 017.00
Median CDS length (bp)	834.00
Median exon length (bp)	162.00
Average exons per gene	5.01
Masked repeat sequence length (Mb)	229.38
Percentage of repeat sequences (%)	50.02

We evaluated the integrity and consistency of the genome sequence as the quality of the genome assembly index has a direct impact on the whole genome quality. The different strategies, such as GC-Depth analysis, SNP analysis, Bench-marking Universal Single-Copy Orthologs (BUSCO), and read alignment using HiFi reads, and ONT reads, Illumina filtered reads, were used to detect plant genome assembly quality (Fig. S2, Fig. S3, and Fig. S4). Firstly, the quality and structure of the genome were confirmed through the structure variation (SV) analysis and mapping of Hi-C data, which revealed no significant occurrences of misalignment intra- or inter-chromosomes. We also used the Illumina paired-end reads, HiFi reads, and ONT reads to map to the *R. banksiae* genome to investigate the completeness of the genome. The *R. banksiae* genome had higher coverages and mapping rates (coverage rates < 99.87%, and mapping rate > 99.34% in Illumina paired-end data, and coverage rates < 99.98% and mapping rate > 99.98% in HiFi and ONT data), suggesting that our genome had almost all of the information provided in these sequenced reads (Table S1 and Table S2). Subsequently, SNP analysis was conducted, which revealed a heterozygous single nucleotide polymorphism (SNP) ratio of 0.337% and a homozygous SNP ratio of 0 in the *R. banksiae* genome. A lower homozygosity rate indicates higher genome accuracy, so our data suggested the *R. banksiae* assembly had a higher degree of base accuracy (Table S3). Finally, we used the BUSCO analysis to assess the completeness of *R. banksiae* genome, revealing 98.30% complete BUSCOs, including 3.50% complete and duplicated BUSCOs and 94.80 % complete and single-copy BUSCOs (Table S4) were observed in our assembly. All these data demonstrate the high accuracy, completeness, and contiguity of the *R. banksiae* genome.

3.2. Genome annotation of R. banksiae

To further gain an accurate gene structure, the RNA-Seq data were used to facilitate reliable genome annotation. In total, we found that 229.38 Mb (50.02%) of the *R. banksiae* genome consisted of repetitive sequences containing 37.59% *de novo* repetitive sequences (Table 1). Among them, 188.57 Mb (41.12%) of the *R. banksiae* genome were annotated as TEs (i.e. DNA, LINE, SINE, and LTR; Fig. 2A; Table S5).

Subsequently, we employed *de novo*, homolog-based, and RNA-Seq methodologies to identify 31, 629 protein-coding genes in the genome of *R. banksiae*, which encompassed 98.5% of the BUSCOs genes found in the embryophyta_odb10 database (Table 2). The completeness of *R. banksiae* was compared with the released *Rosa* genus genomes of *R. rugose*, *R. chinensis*, *R. wichuraiana*, and *R. multiflora*, suggesting that our genome was of high quality and completeness. Among these annotated genes, 30, 872 (97.61%) proteins from *R. rugose* genome

were annotated by using the KOG, NR, TrEMB, Swissprot, KEGG, GO, and InterPro databases (Table S6). The median coding DNA sequence (CDS) and gene lengths were 834.00 bp and 2 017.00 bp (Table 1), respectively. The intron length, exon length, gene length, CDS length, and gene number of the *R. banksiae* genome were compared with other closely species, and found the gene number in *R. banksiae* was smaller than that of other *Rosa* genomes (Fig. S5). Finally, 5 372 non-coding RNAs were annotated, including 655 small nuclear RNAs (snRNAs), 3951 ribosomal RNAs (rRNAs), 632 transfer RNAs (tRNAs), and 134 microRNAs (miRNAs), accounting approximately for 0.86% of the *R. banksiae* genome (Table S7).

Table 2 Genome annotation BUSCO results of *R. banksiae* and other published *Rosa* genus genomes

Statistics	*R. banksiae*	*R. rugose*	*R. chinensis*	*R. wichuraiana*	*R. multiflora*
Complete BUSCOs (%)	98.50	94.40	97.90	93.90	88.40
Fragmented BUSCOs (%)	0.20	0.10	1.00	3.00	4.20
Missing BUSCOs (%)	1.30	5.50	1.10	3.10	7.40

3.3. Genome evolutionary analyses

A total of 528 297 (93.20%) genes were assigned to 40 845 orthogroups by OrthoFinder. In general, G50 refers to the cluster size at which 50% of genes are found in an orthogroup (OG) of that size or larger. O50 represents the minimum number of orthogroups needed to achieve G50 (Emms and Kelly, 2019). Here, 50% were assigned to the largest orthogroups of 7 818 (of which O50 was 7 818) and belonged to the orthogroups with 19 or more genes (of which G50 was 19). A total of 8 475 orthogroups with genes were obtained from all species, of which 223 were entirely single-copy genes. Analysis of 13 species (*P. avium, P. persica, M. domestica, P. communis, P. bretschneideri, P. betulifolia, R. occidentalis, F. vesca, R. banksiae, R. rugosa, R. chinensis, R. wichuraiana*, and *R. multiflora*) revealed 9 676 shared genes, likely constituting the core Rosaceae gene set (Fig. 2, B). In contrast, 19 150 genes were unique to the *R. banksiae* genome (Fig. 2, B). In addition, we identified 28 891, 29 273, and 30 969 collinear gene pairs between *R. banksiae* and *R. wichuraiana*, between *R. banksiae* and *R. rugose*, between *R. banksiae* and *R. chinensis*, accounting for 71.32%, 62.39%, and 61.23% of the total number of genes in these used genomes (Fig. 2, C), respectively.

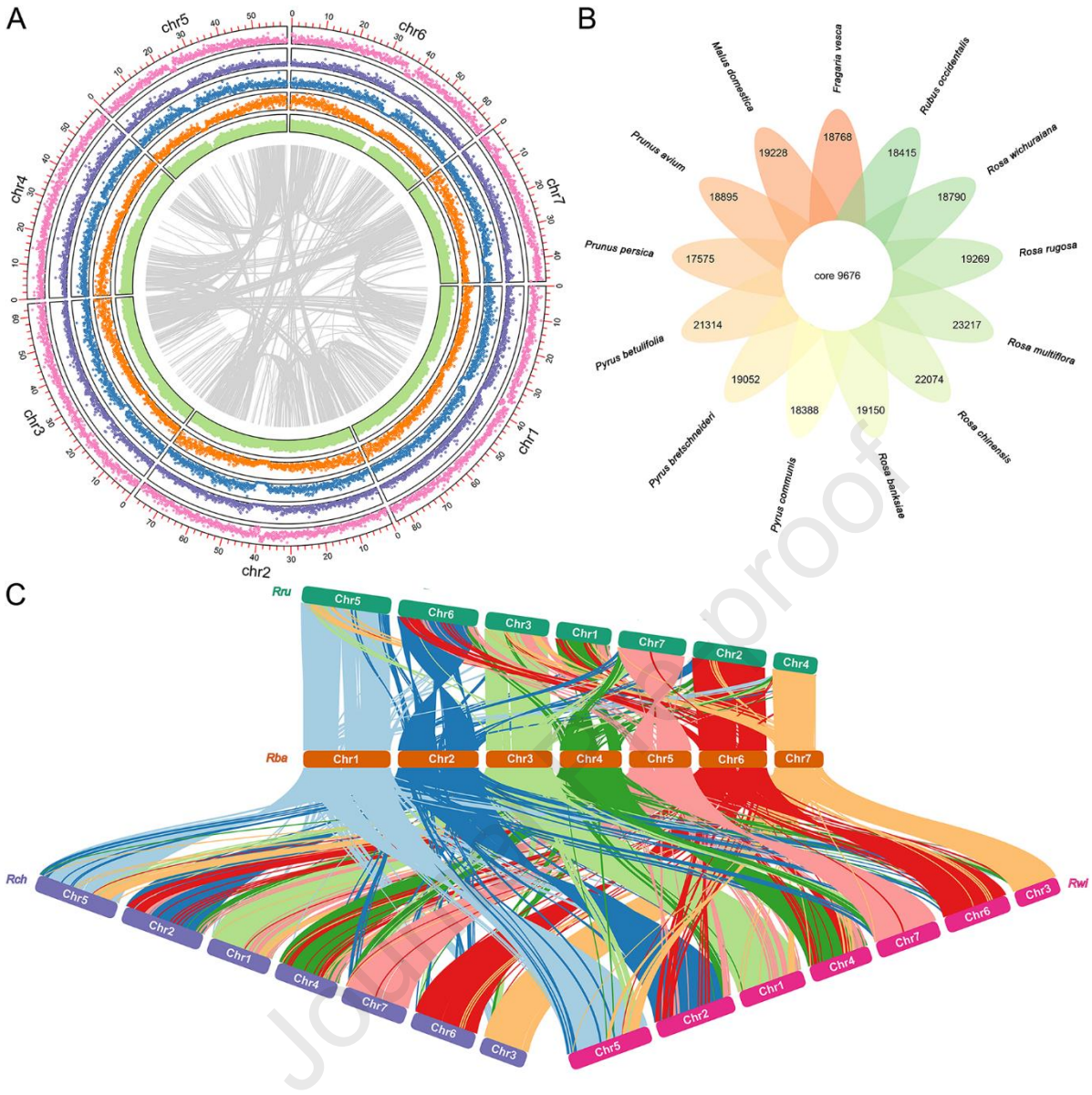


Fig. 2 Genome annotation and comparative genomics analysis of *R. banksiae*

A, The genome annotation of *R. banksiae*. From inner to outer tracks: syntenic relationship, GC content, gene density, DNA-Transposable elements (DNA-TE) density, long interspersed nuclear element (LINE) density, and long terminal repeat (LTR) density. B, The unique and shared gene families were compared among thirteen closely related Rosaceae genomes: *P. avium, P. persica, M. domestica, P. communis, P. bretschneideri, P. betulifolia, R. occidentalis, F. vesca, R. banksiae, R. rugosa, R. chinensis, R. wichuraiana*, and *R. multiflora*. The number of gene families were represented by each number. C, Comparative genomic analysis between *R. rugosa, R. chinensis, R. wichuraiana*, and *R. banksiae.*

Gene family analysis placed 30 924 (97.8%) *R. banksiae* genes into orthogroups, using *Arabidopsis* as an outgroup(Fig. 3). The number of gene families from *R. banksiae* genome that contracted/expanded was prominent at 12 152 gene families compared with that of *R. rugose*. However, the numbers of *R. rugosa* gene families that contracted and expanded were significantly lower -9 977 and 2 534 (*P*-value < 0.05), respectively (Fig. 3, B; Fig. S6). Subsequently, all the genes from the above genomes were clustered into 29, 216 orthogroups, of which 223 single-copy were single-copy orthogroups. These single-copy orthogroups were used to construct a species tree to determine the evolutionary status of *Rosa* (Fig. 3), revealing that *Rosa* is a sister group to *Fragaria*, and their speciation times were approximately 7.04–25.6 Ma. Furthermore, due to the result of recent divergence, *R. banksiae* was more closely related to *R. rugosa* than to other Rosaceae genomes, and the ancestor

of these two genomes split approximately 4.50 Ma (Fig. 3, A). Investigating grapevine (*Vitis vinifera*) and *Rosa* genomes revealed a shared ancestral whole-genome duplication (WGD-γ event), contrasting with the absence of recent WGDs. Comparative analysis identified 19 894 collinear gene pairs across 1 133 syntenic blocks between *R. banksiae* and *V. vinifera*(Fig. S7). Synonymous nucleotide site (Ks) values among *Rosa* genomes align with the WGD-γ event, confirming its eudicot-wide occurrence and excluding recent WGD events (Fig. 3, C and D). This pattern is consistent across various *Rosa* genomes, reinforcing the role of ancient WGDs in plant evolutionary processes.

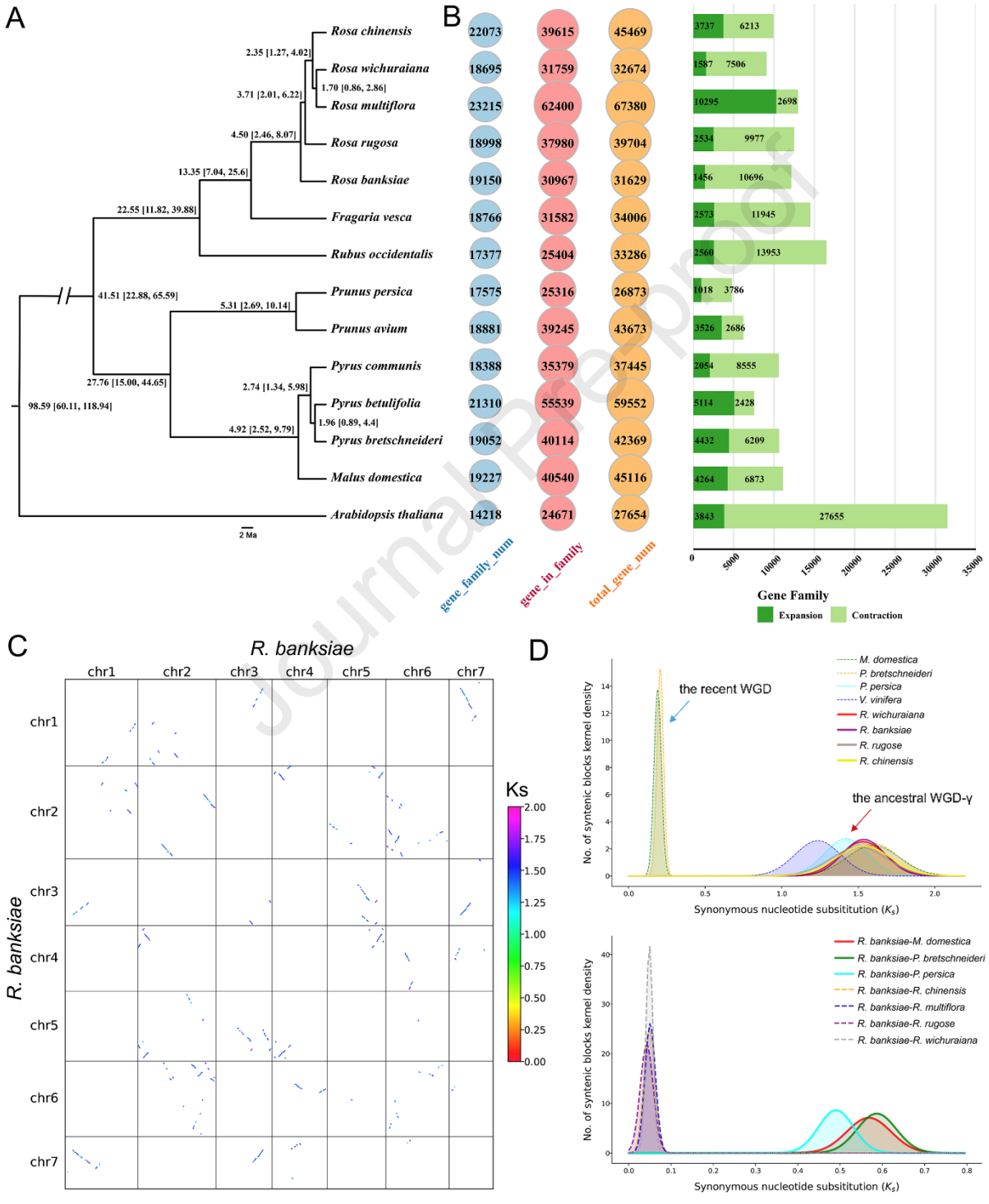


Fig. 3 Phylogenetic analysis and identification of whole-genome duplication (WGD) events

A, The species tree was built according to the 223 single-copy genes of *R. banksiae* and 13 additional specie. The black numbers at each node are divergence

Journal Pre-proof

time to present (in Ma). The length of the line represents the divergence time. B, Numbers of OrthoFinder-identified gene families and CAFE5-calculated numbers of extended or contracted gene families among 14 species. C, The distribution of Ks in *R. banksiae* genome. D, WGD events were indicated by the density distribution of the synonymous substitution rate (Ks) at the intra- and intergenomic levels. The blue and red arrows represented the recent WGD and the ancient WGD-γ, respectively.

3.4. Positively selected genes

Relative to other members of the Rosaceae family, *R. banksiae* is particularly favored by tourists due to its distinctive yellow flowers. To elucidate the genetic foundation underlying the formation of *R. banksiae* flowers, we conducted a search for positively selected genes (PSGs) in its genome. This investigation revealed 36 PSGs in *R. banksiae* (Table S8), in contrast to other genomes within the Rosaceae family. Six genes [(*Rb02g018790.1*: AP2-like ethylene-responsive (*AP2/ERF*); *Rb07g025160.1*: *CONSTITUTIVE PHOTOMORPHOGENIC 1* (*COP1*); *Rb05g014230.1*: ABC transporter (*ABC*); *Rb02g054250.1*: CCCH-type (*C3H*); *Rb05g038170.1*: F-box gene (*F-box*); and *Rb05g021610.1*: RING-H2 finger gene (*RING-H2*)] associated with anthocyanin accumulation were identified as PSGs in *R. banksiae* (Francisco et al., 2013; Kang et al., 2020; Ning et al., 2021; Li et al., 2022a). Among them, the F-box protein notably inhibits COP1 dimerization (Lee et al., 2017), while the RING-H2 protein, by interacting with COP1, regulates anthocyanin accumulation (Kang et al., 2020). Thus, selective pressure on these interacting genes (F-box protein, COP1 and RING-H2) could be instrumental in the evolution of *R. banksiae* flower color, through the modulation of anthocyanin biosynthesis.

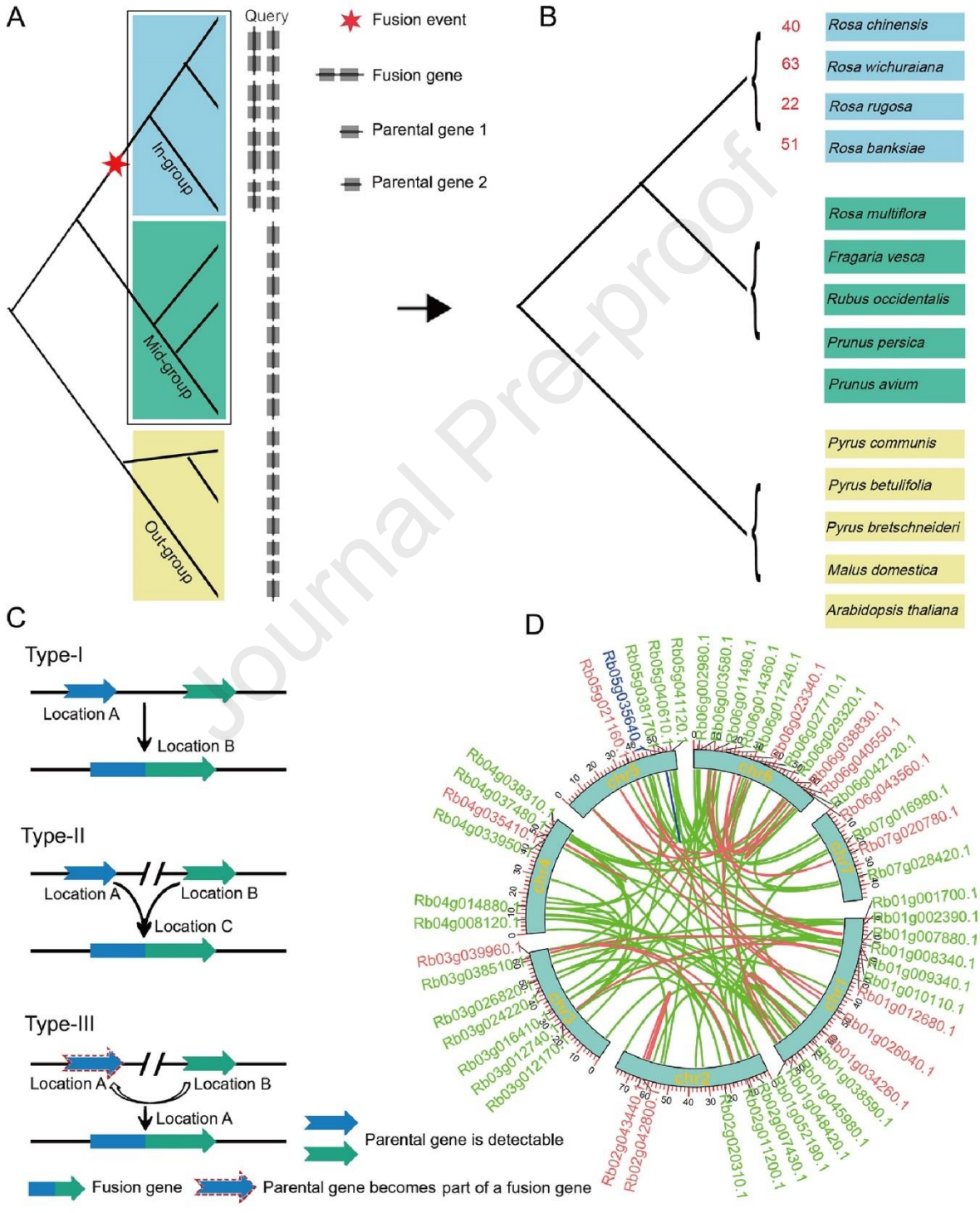
3.5. Gene expression patterns of fusion genes and their parental genes

A gene is determined as a fusion gene if it has at least 2 different parts derived from 2 separate parental genes (Zhou et al., 2022). In single genome studies, it can be difficult to identify between gene fusion events and gene fission events-which occur when a longer homologous parental gene splits into 2 short homologous daughter genes, due to the absence of an outgroup support. Taking into account homologous gene structures and phylogenetic relationships (Fig. 4, A), we identified fusion genes using annotated protein sequences in the Rosaceae genome by GriffinDetector (Zhou et al., 2022). Using the Genome Database for Rosaceae (GDR) annotation dataset for *R. wichuraiana*, *R. rugose*, *R. chinensis*, and *R. banksiae*, and *A. thaliana* as the outgroup species in the phylogeny, 63, 22, 40, and 51 candidate fusion genes were detected in *R. wichuraiana*, *R. rugose*, *R. chinensis*, and *R. banksiae*, respectively (Fig. 4, B; Table S9). Illumina reads are good evidence of the fusion gene. Therefore, we further randomly chose three fusion genes for test. After importing the bam file in IGV-GSAman, we could find reads mapped to the junction area for *Rb05g035640.1, Rb06g023340.1*, and *Rb02g011200.1* (Fig. S8). We compared 181 candidate fusion genes by BLASTP alignment to remove all redundant copies, as some genes in these genomes may be orthologous to each other, as described by Zhou et al. (2022). Finally, of all 176 fusion genes identified, 44 were shared by 2 to 4 genomes (Table S9 and Table 3). However, we found that only 25 shared fusion genes could be properly mapped into the species topology tree (Fig. 3, A, Fig. 4, A and B) when phylogeny was taken into account, indicating that these fusion genes may have first appeared episodically. Therefore, the remaining fusion genes could be classified as species-specific, speculating that they have originated in a more recent evolutionary period (Fig. 3, A).

Table 3 Fusion genes detected in four focus *Rosa* genomes

Shared species	Detail species	Gene numbers	Total gene numbers
Species specific	*R. banksiae*	40	/
	R. rugosa	15	/
	R. wichuraiana	52	/
	R. chinensis	25	/

Shared by four species	*R. banksiae, R. rugosa, R. wichuraiana, R. chinensis*	6	6
Shared by three species	*R. banksiae, R. rugosa, R. chinensis*	6	12
	R. banksiae, R. wichuraiana, R. chinensis	3	
	R. rugosa, R. wichuraiana, R. chinensis	3	
Shared by two species	*R. banksiae, R. rugosa*	2	26
	R. banksiae, R. wichuraiana	6	
	R. rugosa, R. chinensis	4	
	R. wichuraiana, R. chinensis	8	
	R. banksiae, R. chinensis	6	



Journal Pre-proof

Journal Pre-proof

Fig. 4 Fusion event analysis in *Rosa*

A, The process of fusion genes detection, as described by Zhou et al. (2022). B, Identification of fusion genes in the *Rosa* genomes. Species in the blue boxs were focus genomes, fusion genes were identified in all these four genomes for all three different types in Fig. 4, C. C, Origination types of fusion genes, as described by Zhou et al. (2022). D, Circos plot of gene fusions discovered in *R. banksiae*. Red, green, and blue indicated Type-I, Type-II, and Type-III fusion genes.

To determine the possible potential functions of some of the candidate fusion genes, we performed further analysis of the fusion genes to obtain evidence of gene expression in *R. banksiae*. Using RNA-Seq data, we retrieved the values of fragments per kilobase of exon per million fragments (FPKM) for all *R. banksiae* fusion genes and their homologous copies (Table S10). Among the 51 fusion genes, we observed that no detectable FPKM signals were produced at the 6 fusion genes (*Rb01g007880.1*, *Rb01g010110.1*, *Rb02g043440.1*, *Rb03g012740.1*, *Rb06g029320.1*, and *Rb02g020310.1*) identified. However, at their corresponding short homologous copies, these genes produced detectable FPKM signals. Subsequently, among the 45 fusion genes in *R. banksiae* that contained can detect expression signals, four fusion genes (*Rb01g001700.1, Rb01g048420.1, Rb04g014880.1,* and *Rb04g033950.1*) exhibited the same expression patterns as their short homologous copies in all tissues significant *P*-values ($P < 0.05$), 15 fusion genes with significant *P*-values ($P < 0.05$) displayed the same expression patterns as one of their short homologous copies in all tissues, and the remaining fusion genes (26) lacked data for their parental copies or had significantly distinct expression patterns (Table S10 and Table S11). We also calculated the tissue specificity index (Tau, τ), and found that 11 fusion genes were tissue-specific expression (i.e. $\tau \geq 0.9$). However, 8 of these 11 genes are not specifically expressed in the corresponding two parental genes (i.e. $\tau < 0.9$; Table S12). Finally, we examined the gene structure of the fusion gene versus the corresponding parental genes. Among fusion genes in *R. banksiae*, we confirmed that five fusion genes are consistent with the exon-intron structure of the parental genes, *Rb03g026820.1* was shared with *R. chinensis*, while *Rb05g038170.1, Rb03g012170.1, Rb01g052190.1*, and *Rb01g034260.1* were *R. banksiae* specific (Table S9 and Fig. S9), suggesting that most of the fusion genes possess a new ORF (Open Reading Frame) to produce novel functions. Taken together, our findings suggested that many candidate fusion genes have undergone sub-functionalization and neo-functionalization, as suggested by previous research in *Oryza*, *Drosophila,* and human (Rogers et al., 2009; Prince and Pickett, 2002; Zhou et al., 2022).

3.6. The parental gene duplications following fusion events are the dominant structural patterns of origin

To gain better insight into the mechanism of gene fusion formation in the Rosaceae family, the structures of 51 fusion genes were further investigated in *R. banksiae*. First, we divided these identified fusion genes into three types according to the parental gene locus and presence in the genome of the analyzed focus genomes (Wang et al., 2004; Zhou et al., 2022). Type-I fusion genes contain two parental genes that are generated by segmental duplication with distributing in the same region of the chromosome, and then these two parental genes are fused together by mutation or skipping at alternative splice sites (Fig. 4, C). Type-II fusion genes also have two parental genes, formed by duplication, distributed in different regions of the genome, and then they are inserted into adjacent regions for fusion. Type-III fusion genes can only be detected in one parental gene (Fig. 4, C). Here, a duplication event produces two parental genes, one of which is transferred to the neighboring region of the parental gene that did not undergo a duplication event, and then fused together. Therefore, 51 fusion genes were further divided into 13 Type-I, 37 Type-II, and 1 Type-III (Fig. 4, D and Table S13). Notably, we found that only one short homologous copy of 19 fusion genes was detected in *A. thaliana*, but two parental copies were present in other focus species analyzed in the set. These data suggested that some of these parental genes are newly evolved during evolution.

Subsequently, we examined functional constraints by comparing nonsynonymous (Ka) and synonymous (Ks) substitution rates (the ratios of Ka : Ks) between fusion genes and their parental copies using an evolutionary methodology. Generally, it is assumed that the ratio of Ka/Ks for genes that arose after neutrality will be 1, which is higher than unity for genes subject to strong positive selection and lower than unity for genes subject to the functional constraint (Yang, 2007). This approach has been widely and increasingly employed to elucidate gene functionality (Wang et al., 2006). Fig. 5, A exhibited that Type-I and Type-II genes contained different Ka/Ks distributions (Table S14). Purifying selection has a more dominating influence on Type-I genes than Type-II genes, as seen by the tendency of Type-I genes to have higher values of Ka/Ks than Type-II genes. One gene pair (*Rb04g037480.1-Rb05g042140.1/Rb06g044300.1*) had Ka/Ks values larger than 1, indicating that positive selection has mostly affected it (Table S14). These findings suggested that Type-I genes have conserved function and that Type-I genes either have diverging gene expression patterns or may have pleiotropic functions.

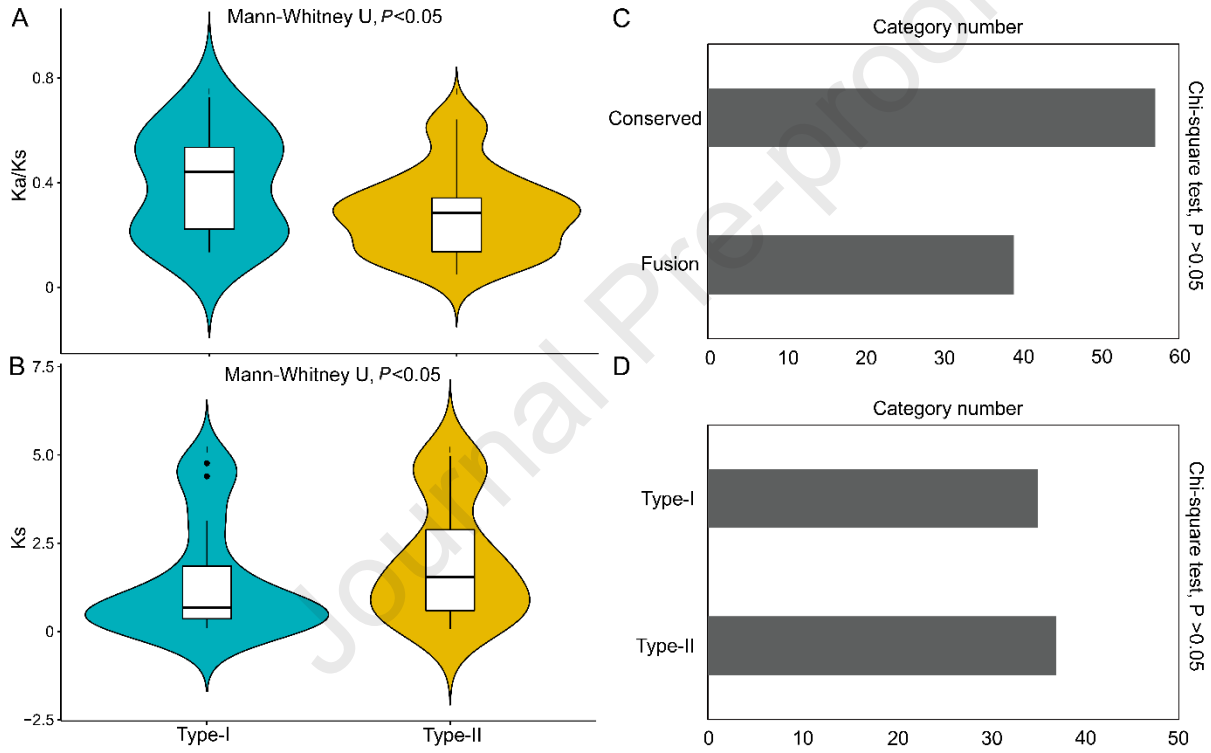


Fig. 5 Comparison of Ka/Ks, Ks, and transcription factor binding sites

A, The selective pressures on fusion genes. B, The Ks distribution of fusion genes. C, Comparison of the number of TFBSs types between conserved and fusion genes. D, Comparison of the number of TFBSs types between Type-I and Type-II fusion genes.

The distribution of Ks values can be used to determine whether the generation of new genes is a continuous process(Wang et al., 2006). In our study, we calculated the Ks values for these identified new fusion genes to determine whether or not the emergence of these genes is a continuous process (Fig. 5, B and Table S14). After excluding those genes with Ks > 2.0, the origin date of fusion genes was estimated using a divergence rate of 1.5×10^{-8} substitutions per synonymous site per year (Koch et al., 2000; Pereira, 2004) and then found that the fusion genes were estimated to have occurred between 3.11 to 66.32 Ma (Table S14). Among them, we found that seven of the fusion genes have Ks values that are less than 0.3, then approximately these genes appeared in the past 10 Ma, indicating that these genes appeared after the differentiation of *Rosa* and *Fragaria* (Fig. 3). Interestingly, the values of Ks in Type-I genes were lower than those in Type-II genes (Fig. 5, B), indicating that

Type-II genes may have originated earlier.

3.7. Fusion genes involved in transcription regulation

We extracted the 2-kb upstream sequences of the start codons (ATG) of full-length fusion genes to predict the *cis*-acting elements using the NSITE website (http://www.softberry.com/berry.phtml) with a *P*-value cutoff of 0.05 (Shahmuradov and Solovyev, 2015). The analysis revealed that all of the fusion genes examined possessed transcription factor binding sites (TFBSs). In total, 39 distinct categories of TFBSs were identified, and on average, each fusion gene contained 18.04 TFBSs (Table S15). Similar to this, 14 975 out of 15 088 conserved genes (99.25%) had TFBSs, and each of these genes contained an average of 73.69 TFBSs (1 103 461 out of 14 975) (Table S15). For conserved genes, we identified 57 categories of TFBSs. The number or category of TFBSs did not differ significantly between fusion genes and conserved genes (chi-square test, P >0.05) (Fig. 5C). Among these identified TFBSs, some unique TFBSs were detected in fusion genes, such as USF, LOB, DST, and S1Fa-like elements, but these TFBSs (such as ARF, B-BOX, and KNOX) were uniquely distributed in conserved genes (Table S15 and Table S16). The number of unique TFBSs identified in this study was lower than that of common TFBSs. Notably, we found a similar pattern for common TFBSs between conserved genes and fusion genes according to the TFBS number for some elements (such as BPC, bZIP, MYB, and WRKY), in addition to unique TFBSs. These distribution patterns of TFBSs between conserved genes and new genes were basically consistent with previous findings(Song et al., 2022), suggesting that new fusion genes have been incorporated into existing regulatory networks. Subsequently, the two main types (i.e. Type-I and Type-II genes) of fusion genes were compared, and found that all Type-I fusion genes and Type-II fusion genes had TFBSs. The average number of TFBSs of Type-I fusion genes was 19.00, and that of Type-II fusion genes was 18.19 (Table S17). Among the number of categories of TFBSs, there was also no difference between the Type-I and Type-II fusion genes (36 versus 37, chi-square test, P > 0.05) (Fig. 5, D). These data suggested that similar TFBSs were randomly distributed between these two types (i.e. Type-I and Type-II genes) of fusion genes.

The gene expression patterns of fusion genes and their corresponding parental genes exhibited gene functional divergence and/or gene expression differences(Zhou et al., 2022), indicating these genes have undergone neo-functionalization and sub-functionalization. TFBSs commonly more affect gene expression differences(Zou et al., 2009; Song et al., 2022). Therefore, we also compared the distributions of TFBSs between fusion genes and their corresponding parental genes, finding that 94.12% of parental 1 genes and 98.00% of parental 2 genes had TFBSs. Notably, about 36.14% of fusion genes with parental 1 genes and about 82.43% of fusion genes with parental 2 genes were determined to contain matching TFBS arrangements (Table S18). Under normal growth conditions, the expression of fusion genes and their corresponding parental genes was conserved, but most of them had mismatched TFBS arrangements. Thus, new fusion genes possibly exhibited divergent expression under specific growth conditions.

3.8. Fusion genes could be involved in the regulation of plant growth and development

Understanding how new genes work may help explain how they first appeared in genomes and then subsequently gained function. In our study, we selected a new gene (i.e., *Rb05g038170.1* has a novel expression pattern with having higher expression level in root and leaf tissues compared to the parental genes; Fig. 6, A) that represented "neo-functionalization" for gene over-expression experiments to further explore whether these expressed fusion genes are associated to adaptive phenotypes. Firstly, we analyzed the gene structure of the fusion gene *Rb05g038170.1* and the parent genes *Rb07g006300.1* and *Rb02g010870.1*, and found that their

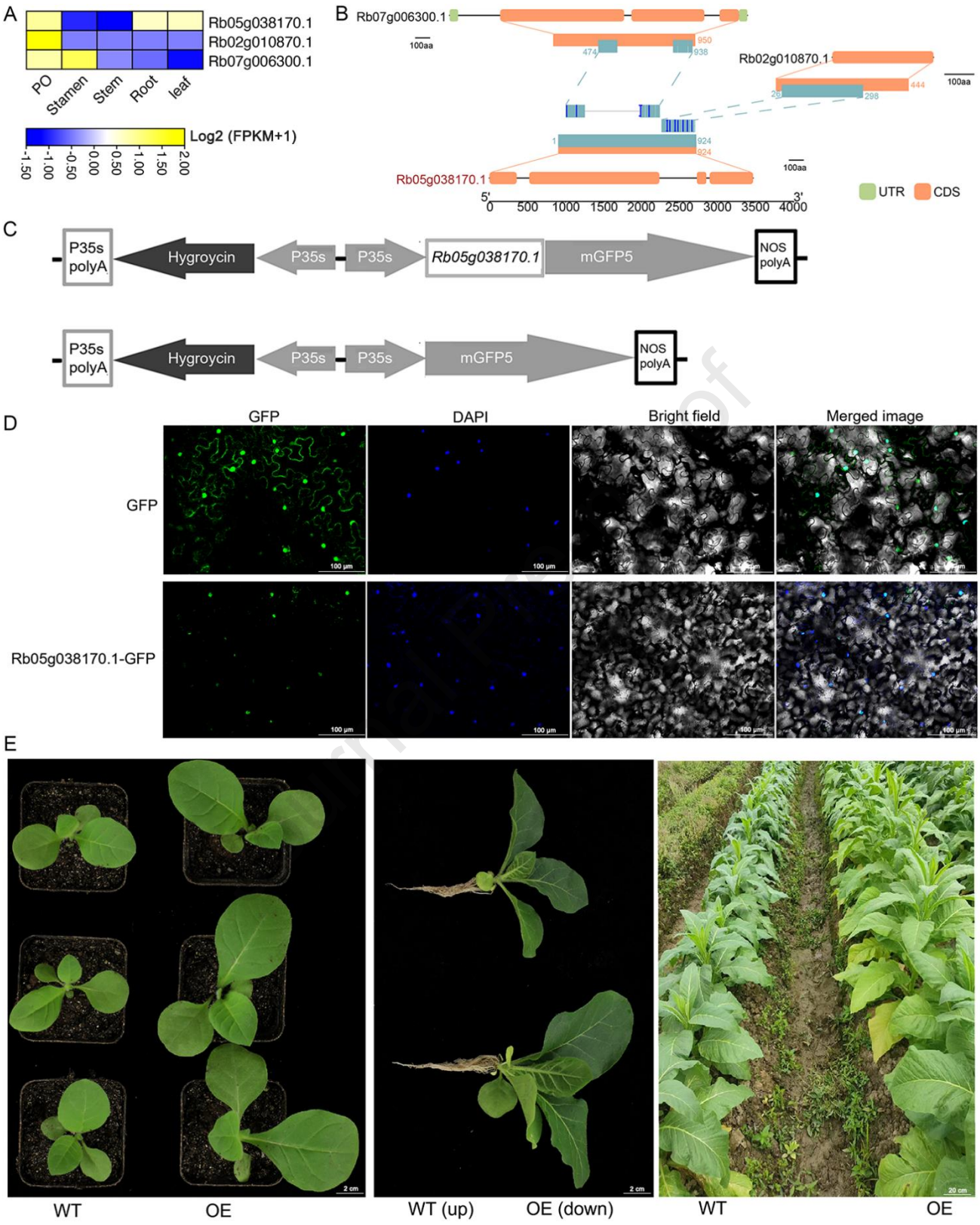


Fig. 6 Functional analysis of *Rb05g038170.1* in *R. banksiae*

A, Expression patterns of the fusion gene *Rb05g038170.1* and the parent genes *Rb07g006300.1* and *Rb02g010870.1*. PO indicated pistil and ovary tissue. B, Structural analysis of the fusion gene *Rb05g038170.1* and the parent genes *Rb07g006300.1* and *Rb02g010870.1*. C, Schematic diagram of pCAMBIA1304-Rb05g038170.1 vector construction. D, Subcellular localization analysis of Rb05g038170.1. The green fluorescent protein (GFP) and fusion constructs were transiently expressed in *N. benthamiana* leaves lower epidermal cells were driven by the cauliflower mosaic virus (CaMV)-35S promoter. 4,6-diamidino-2-phenylindole (DAPI) indicated nuclear localization. E, Phenotypic comparison between wild type (WT) and *Rb05g038170.1* overexpression plants.

structures are basically the same (Fig. 6, B). Then, we constructed a fusion protein 35S: Rb05g038170.1-GFP for transient transformation via *Agrobacterium* infiltration methodology to identify the subcellular location of this protein (Fig. 6, C). The fluorescence signals of the 35S: Rb05g038170.1-GFP protein were discovered to be localized in the nucleus, which was also validated by 4,6-diamidino2-phenylindole (DAPI) staining, whereas fluorescence signals of 35S:GFP were found to be distributed throughout the cell. Taken together, we confirmed this protein Rb05g038170.1 as a nuclear-localized protein that may be capable of performing transcriptional regulatory functions (Fig. 6, D).

Rb05g038170.1 is expressed in almost all tissues in *R. banksiae*, indicating that it may play an important role in plant growth and development. To confirm this role of *Rb05g038170.1*, we constructed an overexpression vector to overexpress this gene in tobacco. The wild-type tobacco was much thinner and shorter than the overexpression tobacco, and this difference is obvious in the early stages of seedling development and only gets more pronounced as the plants develop under identical growing conditions. This result confirmed the role of the new gene in the evolution of *R. banksiae* growth and development (Fig. 6, E).

4. Discussion

ONT sequencing has long-read lengths of 100 kb on average while the most recent generation of HiFi sequencing from PacBio has extremely accurate reads that are 16 kb in length on average, enabling the assembly of intricate and repetitive areas (such as centromeres and heterozygous regions) in the new era of T2T genomics (Miga, 2020; Nurk et al., 2022). Combining these two sequencing technologies, the first human genome at the T2T level is completed in 2022 (Nurk et al., 2022). In plants, some complete or gapless reference genomes, such as watermelon, banana, rice, and *A. thaliana* (Belser et al., 2021; Li et al., 2021; Song et al., 2021; Deng et al., 2022; Wang et al., 2022; Zhang et al., 2022), have also been assembled using these two sequencing technologies. *R. banksiae* is a perennial shrub with significant ornamental and medicinal values. Lack of reference genome hinders genetic improvement and functional genomics of *R. banksiae*. Using a combination of ONT reads, HiFi reads, and Hi-C scaffolding, a high-quality genome for *R. banksiae* was first successfully assembled with contigs N50 = 63.90 Mb and a scaffold N50 = 63.90 Mb. All chromosomes of this genome contained a seven-base telomeric repeat unit (TTTAGGG at the 3′ end or CCCTAAA at the 5′ end) on one end with gap-free, achieving the assembly goals of T2T completeness and high accuracy.

More than 300 species of flowering trees and shrubs, the most of which are deciduous, are included in the *Rosa* genus, which is a member of the Rosaceae family. The majority of the members are diploid (2n = 2x = 16), although few are polyploid. Some sequenced genomes in the genus, such as *R. rugose*, *R. chinensis*, and *R. wichuraiana*(Hibrand Saint-Oyant et al., 2018; Chen et al., 2021a; Zhong et al., 2021), are relatively small, *c.* 2-3 times of the *A. thaliana* genome in size (Michael et al., 2018; Wang et al., 2022). The diversification of the *Rosa* genomes was inferred that occurred between 7.04 and 25.6 Ma, followed by the *Rosa* genus successive split during 2.46-8.07 Ma, which was consistent with results that the *Rosa* genome was separated from the Rosaceae family in the Paleocene, and the genus *Rosa* continued to disintegrate in the Eocene (Baek et al., 2018; Chen et al., 2021a). Phylogenomic analyses suggested that *R. banksiae* and its close relatives in *Rosa* (*R. rugose*, *R. chinensis*, *R. wichuraiana*, and *R. multiflora*) were formed into a monophyletic group, which further supported the findings of previous studies on the evolution of Rosaceae (Hibrand Saint-Oyant et al., 2018; Chen et al., 2021a; Zhong et al., 2021).

The publication of several hundred angiosperm genomes has shown that at least four waves of frequent polyploidization events have contributed to the genomic resources for innovation (van de Peer et al., 2017;

Vanneste et al., 2014; Zhang et al., 2020; Cao et al., 2024). The gene Ks values were calculated in *Malus domestica*, *P. bretschneideri*, *P. persica*, *R. banksiae*, *R. wichuraiana*, *R. rugose*, *R. chinensis,* and *V. vinifera*. According to the Ks values in these used genomes, the genomes of *M. domestica* and *P. bretschneideri* shared peaks at ~0.2, indicative of a recent WGD event, and ~1.6, suggesting an ancient WGD or a eudicot-specific WGT event, respectively (Fawcett et al., 2009; Velasco et al., 2010). However, the other genomes (*P. persica*, *R. banksiae*, *R. wichuraiana*, *R. rugose*, *R. chinensis,* and *V. vinifera*) only contained a single peak at 1.2–1.6, suggesting that the *Rosa* genomes only underwent the eudicot-specific WGT event, similar to *P. persica* and *V. vinifera*. This result was also supported by comparing the whole-genome syntenic analysis, as no recent WGD event were detected in genomes *P. persica*, *R. banksiae*, *R. wichuraiana*, *R. rugose*, *R. chinensis* and *V. vinifera* (Hibrand Saint-Oyant et al., 2018; Chen et al., 2021a; Zhong et al., 2021; Cao et al., 2022).

The study of new genes has been considerably aided by the introduction of phylogenomic techniques and the rapid expansion of high-quality sequenced genomes (Li et al., 2022b). Many studies have demonstrated that fusion genes play a significant role in the origin of new genes in eukaryotes (Kaessmann, 2010; Zhou et al., 2022), which provides a significant driver of adaptive evolutionary novelty. For example, new genes detected in *Phyllostachys edulis* are mainly expressed in rapidly developing shoots, suggesting that they may play an important role in morphogenesis during rapid growth (Jin et al., 2021). Here, we identified 176 fusion genes in the *Rosa* genus (including *R. banksiae*, *R. wichuraiana*, *R. rugose*, *R. chinensis*) at the genome level by GriffinDetector, which diverged very recently. If we assume that these genes have a synonymous substitution rate of 1.5×10^{-8} substitutions per silent site per year (Koch et al., 2000), then these genes arose approximately between 2.55 and 66.32 Ma, which is ten Ma in the primate lineage toward humans 5 times the rate of 0.14 chimeric genes (Marques et al., 2005; Wang et al., 2006).These young parental genes might have originated *de novo* because *A. thaliana* does not have homologous copies that cover more than 20% of the fusion gene. Our findings are in agreement with previous studies that show the high rate of *de novo* gene origination in the genus *Oryza* and that neo-functionalization in gene fusion and fission frequently depends on the duplication of parental genes(Wang and Lan, 2000; Wang et al., 2004; Zhang et al., 2019b; Zhou et al., 2022). Interestingly, compared with Type-II fusion genes, Type-I genes have more relaxed selection pressure and lower Ks values, indicating that these newly evolved Type-I fusion genes may play important roles in driving phenotypic evolution.

Finding out how new fusion genes work may help explain how and why they first originated in the genomes and then developed functions (Wang et al., 2006; Jin et al., 2021; Zhou et al., 2022). Some traits involved in specific regulatory networks can be controlled by new fusion genes (Zhou et al., 2022). Biological variety is produced by the phenotypic change in morphology and development (Zu et al., 2019; Jin et al., 2021). In addition to the more well-known function of gene regulatory systems, a new direction to investigate is whether or not gene evolution contributes to the evolution of genetic basis driving phenotypic change (Chen et al., 2013; Jin et al., 2021). New genes have been recruited into co-expression networks during evolution to play an important role in the fast-growing traits of the shoot in bamboo (Jin et al., 2021). Here, we identified new fusion genes in *Rosa* at the genome level, and then used the overexpression technique to create lines for the fusion gene *Rb05g038170.1*. The significant segregation of phenotypic effects was found between overexpression and wild-type tobacco in growth and development. This is somewhat predicted because prior research has demonstrated that in order for new genes to gain the associated biological roles, they have integrated into and changed the ancestral genetic interaction network(Zu et al., 2019; Jin et al., 2021). Indeed, Huang et al. (2022) identified the new gene *AtEXOV* from *A. thaliana* that evolved important phenotypic effects in morphology and development (Huang et al., 2022). Zhou et al. (2022) identified two fusion genes *Osjap09g15430* and

Osjap07g28390 that they evolved significant phenotypic trait effects in rice, including shoot length, root length, and germination rates (Zhou et al., 2022). These neo-evolutionary interactions provide evidence for phenotypic evolution which suggests that new genes may contribute to the phenotypic evolution of plant morphological traits through neo-functionalization during evolution.

5. Conclusion

In this study, we assembled a gap-free T2T reference genome of *R. banksiae* using HiFi and ONT ultra-long sequencing. The comparative genomic analyses suggested that the *Rosa* genomes experienced only an ancestral WGD-γ event (i.e. eudicot-specific WGD). We first identified new fusion genes in *Rosa* and further divided these genes into three types: Type-I, Type-II, and Type-III. The expression patterns, tissue specificity index, and gene structure analyses suggested that many candidate new fusion genes have undergone neo-functionalization and sub-functionalization. The experiments confirmed that new fusion genes could be able to drive the evolution of plant phenotypes. This high-quality reference genome provided a valuable resource for the study of molecular marker-assisted breeding in *R. banksiae*, and also further revealed novel insights into the genome evolution of *Rosa*.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 32201602, 82304680), the Natural Science Fund of Hubei Province (Grant No. 2023AFB1036), and the Program for Excellent Sci-tech Innovation Teams of Universities in Anhui Province (Grant No. 2022AH010074), and Anhui Provincial Natural Science Foundation (Grant No. 2308085QH295), and Natural Science Research Project of Anhui Educational Committee (Grant No. 2023AH040259), and the Talent Scientific Research Start-up Foundation, Wannan Medical College (Grant No. YR20230110), the Anhui Provincial Department of Education Young Backbone Teachers Overseas Visiting and Training Funding Program (Grant No. JWFX2023033), and Beijing Life Science Academy Project (Grant No. 2023200CC0270).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi: 10.1016/j.hpj.2024.

References

Aganezov, S., Yan, S.M., Soto, D.C., Kirsche, M., Zarate, S., Avdeyev, P., Taylor, D.J., Shafin, K., Shumate, A., Xiao, C., Wagner, J., McDaniel, J., Olson, N.D., Sauria, M.E.G., Vollger, M.R., Rhie, A., Meredith, M., Martin, S., Lee, J., Koren, S., Rosenfeld, J.A., Paten, B., Layer, R., Chin, C.S., Sedlazeck, F.J., Hansen, N.F., Miller, D.E., Phillippy, A.M., Miga, K.H., McCoy, R.C., Dennis, M.Y., Zook, J.M., Schatz, M.C., 2022. A complete reference genome improves analysis of human genetic variation. Science, 376: eabl3533.

Baek, S., Choi, K., Kim, G.-B., Yu, H.-J., Cho, A., Jang, H., Kim, C., Kim, H.-J., Chang, K.S., Kim, J.-H., 2018. Draft genome sequence of wild prunus yedoensis reveals massive inter-specific hybridization between sympatric flowering cherries. Genome Biology, 19: 1-17.

Belser, C., Baurens, F.-C., Noel, B., Martin, G., Cruaud, C., Istace, B., Yahiaoui, N., Labadie, K., Hřibová, E., Doležel, J., 2021. Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing. Communications Biology, 4: 1-12.

Benson, G., 1999. Tandem repeats finder: A program to analyze DNA sequences. Nucleic Acids Res, 27: 573-580.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider, M.,

2003. The swiss-prot protein knowledgebase and its supplement trembl in 2003. Nucleic Acids Res, 31: 365-370.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: A flexible trimmer for illumina sequence data. Bioinformatics, 30: 2114-2120.

Brown, G.G., Formanová, N., Jin, H., Wargachuk, R., Dendy, C., Patil, P., Laforest, M., Zhang, J., Cheung, W.Y., Landry, B.S., 2003. The radish rfo restorer gene of ogura cytoplasmic male sterility encodes a protein with multiple pentatricopeptide repeats. The Plant Journal, 35: 262-272.

Buchfink, B., Xie, C., Huson, D.H., 2015. Fast and sensitive protein alignment using diamond. Nat Methods, 12: 59-60.

Burge, C., Karlin, S., 1997. Prediction of complete gene structures in human genomic DNA. J Mol Biol, 268: 78-94.

Cantarel, B.L., Korf, I., Robb, S.M., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez Alvarado, A., Yandell, M., 2008. Maker: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res, 18: 188-196.

Cao, K., Peng, Z., Zhao, X., Li, Y., Liu, K., Arus, P., Fang, W., Chen, C., Wang, X., Wu, J., 2022. Chromosome-level genome assemblies of four wild peach species provide insights into genome evolution and genetic basis of stress resistance. BMC biology, 20: 1-17.

Cao, Y., Mo, W., Li, Y., Xiong, Y., Wang, H., Zhang, Y., Lin, M., Zhang, L., Li, X., 2024. Functional characterization of nbs-lrr genes reveals an nbs-lrr gene that mediates resistance against fusarium wilt. BMC biology, 22: 45.

Chan, P.P., Lowe, T.M., 2019. Trnascan-se: Searching for trna genes in genomic sequences, in: Gene prediction. Springer, pp. 1-14.

Chen, F., Su, L., Hu, S., Xue, J.-Y., Liu, H., Liu, G., Jiang, Y., Du, J., Qiao, Y., Fan, Y., 2021a. A chromosome-level genome assembly of rugged rose (*Rosa rugosa*) provides insights into its evolution, ecology, and floral characteristics. Horticulture research, 8.

Chen, N., 2004. Using repeat masker to identify repetitive elements in genomic sequences. Current protocols in bioinformatics, 5: 4-10.

Chen, S., Krinsky, B.H., Long, M., 2013. New genes as drivers of phenotypic evolution. Nature Reviews Genetics, 14: 645-660.

Chen, T., Chen, X., Zhang, S., Zhu, J., Tang, B., Wang, A., Dong, L., Zhang, Z., Yu, C., Sun, Y., Chi, L., Chen, H., Zhai, S., Sun, Y., Lan, L., Zhang, X., Xiao, J., Bao, Y., Wang, Y., Zhang, Z., Zhao, W., 2021b. The genome sequence archive family: Toward explosive data growth and diverse data types. Genomics Proteomics Bioinformatics, 19: 578-583.

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., Li, H., 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods, 18: 170-175.

Deng, Y., Liu, S., Zhang, Y., Tan, J., Li, X., Chu, X., Xu, B., Tian, Y., Sun, Y., Li, B., 2022. A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provide important resources for gene discovery and breeding. Molecular Plant, 15: 1268-1284.

Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., Aiden, E.L., 2017. De novo assembly of the aedes aegypti genome using hi-c yields chromosome-length scaffolds. Science, 356: 92-95.

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S., Huntley, M.H., Lander, E.S., Aiden, E.L., 2016. Juicer provides a one-click system for analyzing loop-resolution hi-c experiments. Cell Syst, 3: 95-98.

Elrouby, N., Bureau, T.E., 2010. Bs1, a new chimeric gene formed by retrotransposon-mediated exon shuffling in maize. Plant Physiol, 153: 1413-1424.

Emms, D.M., Kelly, S., 2019. Orthofinder: Phylogenetic orthology inference for comparative genomics. Genome Biol, 20: 238.

Fawcett, J.A., Maere, S., Van De Peer, Y., 2009. Plants with double genomes might have had a better chance to survive the cretaceous–tertiary extinction event. Proceedings of the National Academy of Sciences, 106: 5737-5742.

Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., Smit, A.F., 2020. Repeatmodeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A, 117: 9451-9457.

Francisco, R.M., Regalado, A., Ageorges, A., Burla, B.J., Bassin, B., Eisenach, C., Zarrouk, O., Vialet, S., Marlin, T., Chaves, M.M., Martinoia, E., Nagy, R., 2013. Abcc1, an atp binding cassette protein from grape berry, transports anthocyanidin 3-o-glucosides. Plant Cell, 25: 1840-1854.

Gene Ontology, C., 2019. The gene ontology resource: 20 years and still going strong. Nucleic acids research, 47: D330-D338.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K., Jr., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., Salzberg, S.L., White, O., 2003. Improving the arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res, 31: 5654-5666.

Han, M.V., Thomas, G.W., Lugo-Martinez, J., Hahn, M.W., 2013. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using cafe 3. Mol Biol Evol, 30: 1987-1997.

Hibrand Saint-Oyant, L., Ruttink, T., Hamama, L., Kirov, I., Lakhwani, D., Zhou, N.-N., Bourke, P.M., Daccord, N., Leus, L., Schulz, D., 2018. A high-quality genome sequence of rosa chinensis to elucidate ornamental traits. Nature plants, 4: 473-484.

Huang, Y., Chen, J., Dong, C., Sosa, D., Xia, S., Ouyang, Y., Fan, C., Li, D., Mortola, E., Long, M., 2022. Species-specific partial gene duplication in arabidopsis thaliana evolved novel phenotypic effects on morphological traits under strong positive selection. The Plant Cell, 34: 802-817.

Jain, C., Rhie, A., Zhang, H., Chu, C., Walenz, B.P., Koren, S., Phillippy, A.M., 2020. Weighted minimizer sampling improves long read mapping. Bioinformatics, 36:

i111-i118.

Jiang, L., Lin, M., Wang, H., Song, H., Zhang, L., Huang, Q., Chen, R., Song, C., Li, G., Cao, Y., 2022. Haplotype-resolved genome assembly of *Bletilla striata* (thunb.) reichb.F. To elucidate medicinal value. Plant J, 111: 1340-1353.

Jiang, T., Liu, Y., Jiang, Y., Li, J., Gao, Y., Cui, Z., Liu, Y., Liu, B., Wang, Y., 2020. Long-read-based human genomic structural variation detection with cutesv. Genome Biol, 21: 189.

Jin, G., Ma, P.-F., Wu, X., Gu, L., Long, M., Zhang, C., Li, D.-Z., 2021. New genes interacted with recent whole-genome duplicates in the fast stem growth of bamboos. Molecular biology and evolution, 38: 5752-5768.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.Y., Lopez, R., Hunter, S., 2014. Interproscan 5: Genome-scale protein function classification. Bioinformatics, 30: 1236-1240.

Kaessmann, H., 2010. Origins, evolution, and phenotypic impact of new genes. Genome Research, 20: 1313-1326.

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R., Bateman, A., Finn, R.D., Petrov, A.I., 2018. Rfam 13.0: Shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res, 46: D335-D342.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., Yamanishi, Y., 2008. Kegg for linking genomes to life and the environment. Nucleic Acids Res, 36: D480-484.

Kang, H., Zhang, T.T., Fu, L.L., You, C.X., Wang, X.F., Hao, Y.J., 2020. The apple ring-h2 protein mdcip8 regulates anthocyanin accumulation and hypocotyl elongation by interacting with mdcop1. Plant Sci, 301: 110665.

Kim, D., Paggi, J.M., Park, C., Bennett, C., Salzberg, S.L., 2019. Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. Nat Biotechnol, 37: 907-915.

Koch, M.A., Haubold, B., Mitchell-Olds, T., 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in arabidopsis, arabis, and related genera (Brassicaceae). Molecular Biology and Evolution, 17: 1483-1498.

Kumar, S., Stecher, G., Suleski, M., Hedges, S.B., 2017. Timetree: A resource for timelines, timetrees, and divergence times. Mol Biol Evol, 34: 1812-1819.

Lagesen, K., Hallin, P., Rodland, E.A., Staerfeldt, H.H., Rognes, T., Ussery, D.W., 2007. Rnammer: Consistent and rapid annotation of ribosomal rna genes. Nucleic Acids Res, 35: 3100-3108.

Lee, B.D., Kim, M.R., Kang, M.Y., Cha, J.Y., Han, S.H., Nawkar, G.M., Sakuraba, Y., Lee, S.Y., Imaizumi, T., McClung, C.R., Kim, W.Y., Paek, N.C., 2017. The f-box protein fkf1 inhibits dimerization of cop1 in the control of photoperiodic flowering. Nat Commun, 8: 2259.

Leus, L., Laere, K.V., Riek, J.D., Huylenbroeck, J.V., 2018. Rose, in: Ornamental crops. Springer, pp. 719-767.

Li, D., Yang, J., Pak, S., Zeng, M., Sun, J., Yu, S., He, Y., Li, C., 2022a. Puc3h35 confers drought tolerance by enhancing lignin and proanthocyanidin biosynthesis in the roots of populus ussuriensis. New Phytol, 233: 390-408.

Li, J., Shen, J., Wang, R., Chen, Y., Zhang, T., Wang, H., Guo, C., Qi, J., 2022b. The nearly complete assembly of the cercis chinensis genome and fabaceae phylogenomic studies provide insights into new gene evolution. Plant Communications, 100422.

Li, K., Jiang, W., Hui, Y., Kong, M., Feng, L.-Y., Gao, L.-Z., Li, P., Lu, S., 2021. Gapless indica rice genome reveals synergistic contributions of active transposable elements and segmental duplications to rice genome evolution. Molecular Plant, 14: 1745-1756.

Long, M., Betran, E., Thornton, K., Wang, W., 2003. The origin of new genes: Glimpses from the young and old. Nat Rev Genet, 4: 865-875.

Long, M., VanKuren, N.W., Chen, S., Vibranovski, M.D., 2013. New gene evolution: Little did we know. Annual Review of Genetics, 47: 307.

Longhi, S., Giongo, L., Buti, M., Surbanovski, N., Viola, R., Velasco, R., Ward, J.A., Sargent, D.J., 2014. Molecular genetics and genomics of the rosoideae: State of the art and future perspectives. Hortic Res, 1: 1.

Macías, L.G., Barrio, E., Toft, C., 2020. Gwidecodeml: A python package for testing evolutionary hypotheses at the genome-wide level. G3: Genes, Genomes, Genetics, 10: 4369-4372.

Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R.N., Potter, S.C., Finn, R.D., Lopez, R., 2019. The embl-ebi search and sequence analysis tools apis in 2019. Nucleic Acids Res, 47: W636-W641.

Majoros, W.H., Pertea, M., Salzberg, S.L., 2004. Tigrscan and glimmerhmm: Two open source ab initio eukaryotic gene-finders. Bioinformatics, 20: 2878-2879.

Marques, A.C., Dupanloup, I., Vinckenbosch, N., Reymond, A., Kaessmann, H., 2005. Emergence of young human genes after a burst of retroposition in primates. PLoS Biology, 3: e357.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A., 2010. The genome analysis toolkit: A mapreduce framework for analyzing next-generation DNA sequencing data. Genome Res, 20: 1297-1303.

Mehta, T.K., Koch, C., Nash, W., Knaack, S.A., Sudhakar, P., Olbei, M., Bastkowski, S., Penso-Dolfin, L., Korcsmaros, T., Haerty, W., Roy, S., Di-Palma, F., 2021.

Evolution of regulatory networks associated with traits under selection in cichlids. Genome Biol, 22: 25.

Members, C.-N., 2023. Database resources of the national genomics data center, china national center for bioinformation in 2023. Nucleic Acids Research, 51: D18.

Michael, T.P., Jupe, F., Bemm, F., Motley, S.T., Sandoval, J.P., Lanz, C., Loudet, O., Weigel, D., Ecker, J.R., 2018. High contiguity arabidopsis thaliana genome assembly with a single nanopore flow cell. Nat Commun, 9: 541.

Miga, K.H., 2020. Centromere studies in the era of 'telomere-to-telomere'genomics. Experimental Cell Research, 394: 112127.

Miga, K.H., Koren, S., Rhie, A., Vollger, M.R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G.A., Schneider, V.A., Potapova, T., Wood, J., Chow, W., Armstrong, J., Fredrickson, J., Pak, E., Tigyi, K., Kremitzki, M., Markovic, C., Maduro, V., Dutra, A., Bouffard, G.G., Chang, A.M., Hansen, N.F., Wilfert, A.B., Thibaud-Nissen, F., Schmitt, A.D., Belton, J.M., Selvaraj, S., Dennis, M.Y., Soto, D.C., Sahasrabudhe, R., Kaya, G., Quick, J., Loman, N.J., Holmes, N., Loose, M., Surti, U., Risques, R.A., Graves Lindsay, T.A., Fulton, R., Hall, I., Paten, B., Howe, K., Timp, W., Young, A., Mullikin, J.C., Pevzner, P.A., Gerton, J.L., Sullivan, B.A., Eichler, E.E., Phillippy, A.M., 2020. Telomere-to-telomere assembly of a complete human x chromosome. Nature, 585: 79-84.

Mikheenko, A., Bzikadze, A.V., Gurevich, A., Miga, K.H., Pevzner, P.A., 2019. Tandemmapper and tandemquast: Mapping long reads and assessing/improving assembly quality in extra-long tandem repeats. BioRxiv.

Natarajan, P., Murugesan, A.K., Govindan, G., Gopalakrishnan, A., Kumar, R., Duraisamy, P., Balaji, R., Tanuja, Shyamli, P.S., Parida, A.K., Parani, M., 2021. A reference-grade genome identifies salt-tolerance genes from the salt-secreting mangrove species avicennia marina. Commun Biol, 4: 851.

Navratilova, P., Toegelova, H., Tulpova, Z., Kuo, Y.T., Stein, N., Dolezel, J., Houben, A., Simkova, H., Mascher, M., 2022. Prospects of telomere-to-telomere assembly in barley: Analysis of sequence gaps in the morexv3 reference genome. Plant Biotechnol J, 20: 1373-1386.

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. Mol Biol Evol, 32: 268-274.

Ning, Z., Hu, K., Zhou, Z., Zhao, D., Tang, J., Wang, H., Li, L., Ding, C., Chen, X., Yao, G., Zhang, H., 2021. Iberf71, with ibmyb340 and ibbhlh2, coregulates anthocyanin accumulation by binding to the ibans1 promoter in purple-fleshed sweet potato (*Ipomoea batatas* L.). Plant Cell Rep, 40: 157-169.

Nowak, M.D., Birkeland, S., Mandáková, T., Roy Choudhury, R., Guo, X., Gustafsson, A.L.S., Gizaw, A., Schrøder-Nielsen, A., Fracassetti, M., Brysting, A.K., 2021. The genome of draba nivalis shows signatures of adaptation to the extreme environmental stresses of the arctic. Molecular Ecology Resources, 21: 661-676.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., 2022. The complete sequence of a human genome. Science, 376: 44-53.

Pereira, V., 2004. Insertion bias and purifying selection of retrotransposons in the arabidopsis thalianagenome. Genome biology, 5: 1-10.

Pertea, M., Kim, D., Pertea, G.M., Leek, J.T., Salzberg, S.L., 2016. Transcript-level expression analysis of rna-seq experiments with hisat, stringtie and ballgown. Nat Protoc, 11: 1650-1667.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., Salzberg, S.L., 2015. Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. Nat Biotechnol, 33: 290-295.

Price, A.L., Jones, N.C., Pevzner, P.A., 2005. De novo identification of repeat families in large genomes. Bioinformatics, 21: i351-358.

Prince, V.E., Pickett, F.B., 2002. Splitting pairs: The diverging fates of duplicated genes. Nature Reviews Genetics, 3: 827-837.

Raymond, O., Gouzy, J., Just, J., Badouin, H., Verdenaud, M., Lemainque, A., Vergne, P., Moja, S., Choisne, N., Pont, C., 2018. The rosa genome provides new insights into the domestication of modern roses. Nature Genetics, 50: 772-777.

Robinson, J.T., Turner, D., Durand, N.C., Thorvaldsdottir, H., Mesirov, J.P., Aiden, E.L., 2018. Juicebox.Js provides a cloud-based visualization system for hi-c data. Cell Syst, 6: 256-258 e251.

Rogers, R.L., Bedford, T., Hartl, D.L., 2009. Formation and longevity of chimeric and duplicate genes in drosophila melanogaster. Genetics, 181: 313-322.

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E., Dekker, J., Barillot, E., 2015. Hic-pro: An optimized and flexible pipeline for hi-c data processing. Genome Biol, 16: 259.

Shahmuradov, I.A., Solovyev, V.V., 2015. Nsite, nsiteh and nsitem computer tools for studying transcription regulatory elements. Bioinformatics, 31: 3544-3545.

Shen, Z., Li, W., Li, Y., Liu, M., Cao, H., Provart, N., Ding, X., Sun, M., Tang, Z., Yue, C., 2021. The red flower wintersweet genome provides insights into the evolution of magnoliids and the molecular mechanism for tepal color development. The Plant Journal: for Cell and Molecular Biology.

Shulaev, V., Sargent, D.J., Crowhurst, R.N., Mockler, T.C., Folkerts, O., Delcher, A.L., Jaiswal, P., Mockaitis, K., Liston, A., Mane, S.P., Burns, P., Davis, T.M., Slovin, J.P., Bassil, N., Hellens, R.P., Evans, C., Harkins, T., Kodira, C., Desany, B., Crasta, O.R., Jensen, R.V., Allan, A.C., Michael, T.P., Setubal, J.C., Celton, J.M., Rees, D.J., Williams, K.P., Holt, S.H., Ruiz Rojas, J.J., Chatterjee, M., Liu, B., Silva, H., Meisel, L., Adato, A., Filichkin, S.A., Troggio, M., Viola, R., Ashman, T.L., Wang,

H., Dharmawardhana, P., Elser, J., Raja, R., Priest, H.D., Bryant, D.W., Jr., Fox, S.E., Givan, S.A., Wilhelm, L.J., Naithani, S., Christoffels, A., Salama, D.Y., Carter, J., Lopez Girona, E., Zdepski, A., Wang, W., Kerstetter, R.A., Schwab, W., Korban, S.S., Davik, J., Monfort, A., Denoyes-Rothan, B., Arus, P., Mittler, R., Flinn, B., Aharoni, A., Bennetzen, J.L., Salzberg, S.L., Dickerman, A.W., Velasco, R., Borodovsky, M., Veilleux, R.E., Folta, K.M., 2011. The genome of woodland strawberry (*Fragaria vesca*). Nat Genet, 43: 109-116.

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.

Smolka, M., Paulin, L.F., Grochowski, C.M., Horner, D.W., Mahmoud, M., Behera, S., Kalef-Ezra, E., Gandhi, M., Hong, K., Pehlivan, D., 2022. Comprehensive structural variant detection: From mosaic to population-level. BioRxiv, 2022-2004.

Song, H., Guo, Z., Zhang, X., Sui, J., 2022. De novo genes in *Arachis hypogaea* cv. Tifrunner: Systematic identification, molecular evolution, and potential contributions to cultivated peanut. Plant J, 111: 1081-1095.

Song, J.M., Xie, W.Z., Wang, S., Guo, Y.X., Koo, D.H., Kudrna, D., Gong, C., Huang, Y., Feng, J.W., Zhang, W., 2021. Two gap-free reference genomes and a global view of the centromere architecture in rice. Molecular Plant, 14: 1757-1767.

Stamatakis, A., 2014. Raxml version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30: 1312-1313.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., Morgenstern, B., 2006. Augustus: Ab initio prediction of alternative transcripts. Nucleic Acids Res, 34: W435-439.

Sun, P., Jiao, B., Yang, Y., Shan, L., Li, T., Li, X., Xi, Z., Wang, X., Liu, J., 2022. Wgdi: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. Mol Plant, 15: 1841-1851.

Van de Peer, Y., Mizrachi, E., Marchal, K., 2017. The evolutionary significance of polyploidy. Nature Reviews Genetics, 18: 411-424.

Van Huylenbroeck, J., 2018. Ornamental crops. Springer.

van Rengs, W.M.J., Schmidt, M.H., Effgen, S., Le, D.B., Wang, Y., Zaidan, M., Huettel, B., Schouten, H.J., Usadel, B., Underwood, C.J., 2022. A chromosome scale tomato genome built from complementary pacbio and nanopore sequences alone reveals extensive linkage drag during breeding. Plant J, 110: 572-588.

Vanneste, K., Baele, G., Maere, S., Van de Peer, Y., 2014. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the cretaceous–paleogene boundary. Genome research, 24: 1334-1347.

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P., Bhatnagar, S.K., Troggio, M., Pruss, D., 2010. The genome of the domesticated apple (*Malus* × *domestica* borkh.). Nature genetics, 42: 833-839.

Wang, B., Yang, X., Jia, Y., Xu, Y., Jia, P., Dang, N., Wang, S., Xu, T., Zhao, X., Gao, S., 2022. High-quality arabidopsis thaliana genome assembly with nanopore and hifi long reads. Genomics, Proteomics & Bioinformatics, 20: 4-13.

Wang, J., Yang, Z., Lei, T., Zhang, Y., Xiao, Q., Yu, Z., Zhang, J., Sun, S., Xu, Q., Shen, S., 2023. A likely autotetraploidization event shaped the chinese mahogany (*Toona sinensis*) genome. Horticultural Plant Journal, 9: 306-320.

Wang, W., Lan, H., 2000. Rapid and parallel chromosomal number reductions in muntjac deer inferred from mitochondrial DNA phylogeny. Molecular biology and evolution, 17: 1326-1333.

Wang, W., Yu, H., Long, M., 2004. Duplication-degeneration as a mechanism of gene fission and the origin of new genes in drosophila species. Nature genetics, 36: 523-527.

Wang, W., Zheng, H., Fan, C., Li, J., Shi, J., Cai, Z., Zhang, G., Liu, D., Zhang, J., Vang, S., 2006. High rate of chimeric gene origination by retroposition in plant genomes. The Plant Cell, 18: 1791-1802.

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., Topfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.S., Phillippy, A.M., Schatz, M.C., Myers, G., DePristo, M.A., Ruan, J., Marschall, T., Sedlazeck, F.J., Zook, J.M., Li, H., Koren, S., Carroll, A., Rank, D.R., Hunkapiller, M.W., 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol, 37: 1155-1162.

Wu, J., Wang, Z., Shi, Z., Zhang, S., Ming, R., Zhu, S., Khan, M.A., Tao, S., Korban, S.S., Wang, H., Chen, N.J., Nishio, T., Xu, X., Cong, L., Qi, K., Huang, X., Wang, Y., Zhao, X., Wu, J., Deng, C., Gou, C., Zhou, W., Yin, H., Qin, G., Sha, Y., Tao, Y., Chen, H., Yang, Y., Song, Y., Zhan, D., Wang, J., Li, L., Dai, M., Gu, C., Wang, Y., Shi, D., Wang, X., Zhang, H., Zeng, L., Zheng, D., Wang, C., Chen, M., Wang, G., Xie, L., Sovero, V., Sha, S., Huang, W., Zhang, S., Zhang, M., Sun, J., Xu, L., Li, Y., Liu, X., Li, Q., Shen, J., Wang, J., Paull, R.E., Bennetzen, J.L., Wang, J., Zhang, S., 2013. The genome of the pear (*Pyrus bretschneideri* Rehd.). Genome Res, 23: 396-408.

Xin, H., Ji, F., Wu, J., Zhang, S., Yi, C., Zhao, S., Cong, R., Zhao, L., Zhang, H., Zhang, Z., 2023. Chromosome-scale genome assembly of marigold (*Tagetes erecta* L.): An ornamental plant and feedstock for industrial lutein production. Horticultural Plant Journal, 9: 1119-1130.

Journal Pre-proof

Xu, M., Guo, L., Gu, S., Wang, O., Zhang, R., Peters, B.A., Fan, G., Liu, X., Xu, X., Deng, L., Zhang, Y., 2020. Tgs-gapcloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. Gigascience, 9: giaa094.

Xu, Z., Wang, H., 2007. Ltr_finder: An efficient tool for the prediction of full-length ltr retrotransposons. Nucleic Acids Res, 35: W265-268.

Yang, Z., 2007. Paml 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol, 24: 1586-1591.

Yoshioka, K., Moeder, W., Kang, H.G., Kachroo, P., Masmoudi, K., Berkowitz, G., Klessig, D.F., 2006. The chimeric arabidopsis cyclic nucleotide-gated ion channel11/12 activates multiple pathogen resistance responses. Plant Cell, 18: 747-763.

Zhang, L., Hu, J., Han, X., Li, J., Gao, Y., Richards, C.M., Zhang, C., Tian, Y., Liu, G., Gul, H., Wang, D., Tian, Y., Yang, C., Meng, M., Yuan, G., Kang, G., Wu, Y., Wang, K., Zhang, H., Wang, D., Cong, P., 2019a. A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. Nat Commun, 10: 1494.

Zhang, L., Ren, Y., Yang, T., Li, G., Chen, J., Gschwend, A.R., Yu, Y., Hou, G., Zi, J., Zhou, R., 2019b. Rapid evolution of protein diversity by de novo origination in oryza. Nature Ecology & Evolution, 3: 679-690.

Zhang, L., Wu, S., Chang, X., Wang, X., Zhao, Y., Xia, Y., Trigiano, R.N., Jiao, Y., Chen, F., 2020. The ancient wave of polyploidization events in flowering plants and their facilitated adaptation to environmental stress. Plant, Cell & Environment, 43: 2847-2856.

Zhang, Y., Fu, J., Wang, K., Han, X., Yan, T., Su, Y., Li, Y., Lin, Z., Qin, P., Fu, C., 2022. The telomere-to-telomere gap-free genome of four rice parents reveals sv and pav patterns in hybrid rice breeding. Plant Biotechnology Journal, 20: 1642-1644.

Zheng, T., Li, P., Zhuo, X., Liu, W., Qiu, L., Li, L., Yuan, C., Sun, L., Zhang, Z., Wang, J., 2022. The chromosome-level genome provides insight into the molecular mechanism underlying the tortuous-branch phenotype of prunus mume. New Phytologist, 235: 141-156.

Zhong, M.C., Jiang, X.D., Yang, G.Q., Cui, W.H., Suo, Z.Q., Wang, W.J., Sun, Y.B., Wang, D., Cheng, X.C., Li, X.M., 2021. Rose without prickle: Genomic insights linked to moisture adaptation. National science review, 8: nwab092.

Zhou, C., Zhu, C., Tian, C., Xie, S., Xu, K., Huang, L., Wen, S., Zhang, C., Lai, Z., Lin, Y., 2023. The chromosome-scale genome assembly of *Jasminum sambac* var. *unifoliatum* provides insights into the formation of floral fragrance. Horticultural Plant Journal, 9: 1131-1148.

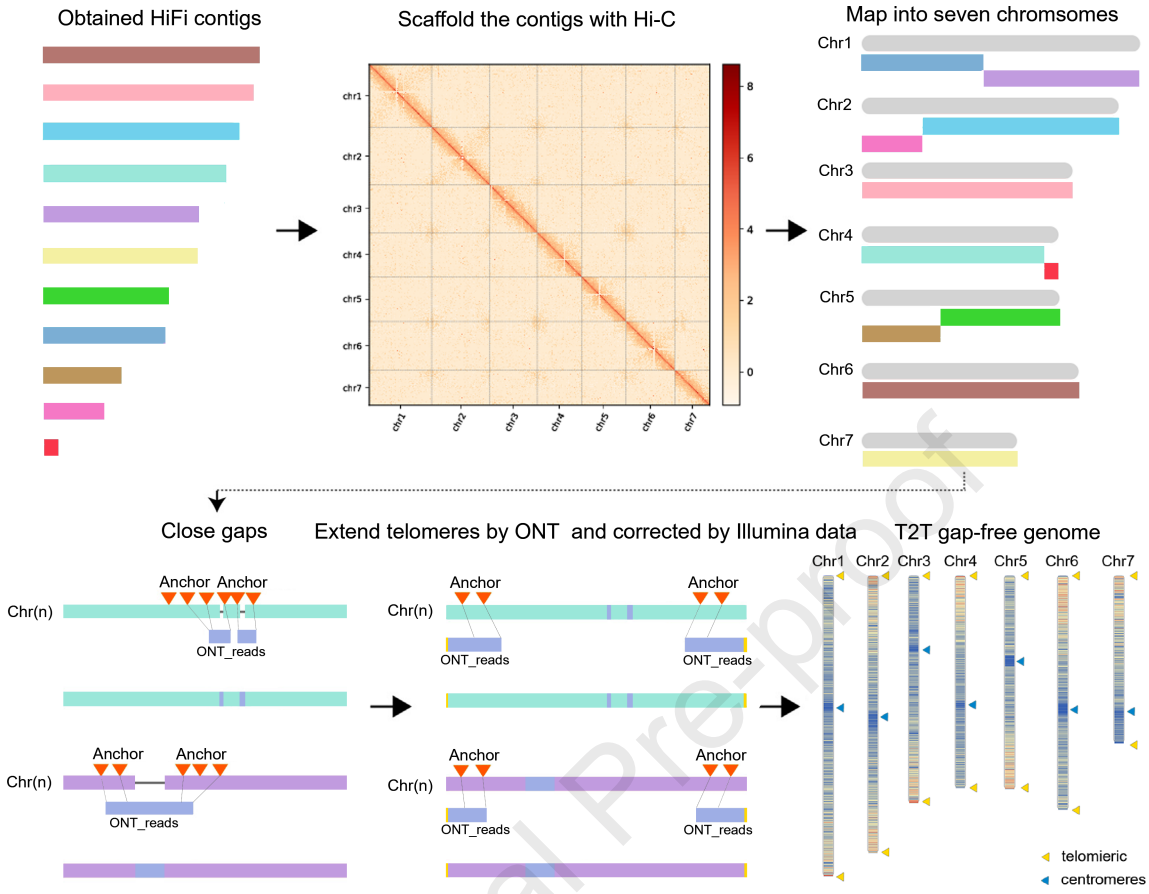
Zhou, Y., Zhang, C., Zhang, L., Ye, Q., Liu, N., Wang, M., Long, G., Fan, W., Long, M., Wing, R.A., 2022. Gene fusion as an important mechanism to generate new genes in the genus *Oryza*. Genome Biology, 23: 1-23.

Zhuang, D., Ma, C., Xue, L., Li, Z., Wang, C., Lei, J., Yuan, X., 2021. Transcriptome and de novo analysis of *Rosa xanthina* f. *spontanea* in response to cold stress. BMC Plant Biology, 21: 1-11.

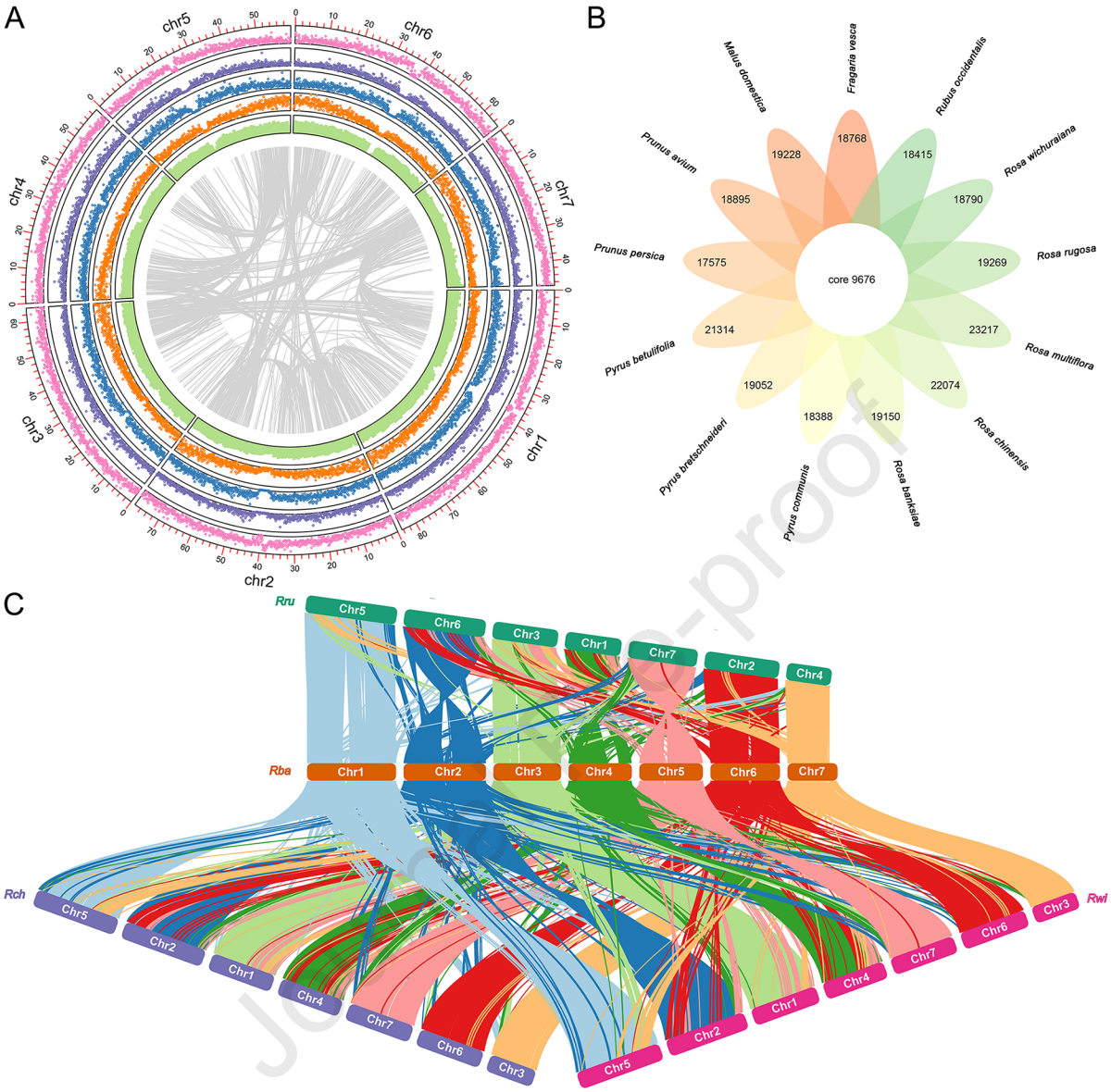
Zou, C., Lehti-Shiu, M.D., Thomashow, M., Shiu, S.-H., 2009. Evolution of stress-regulated gene expression in duplicate genes of *Arabidopsis thaliana.* PLoS Genetics, 5: e1000581.

Zu, J., Gu, Y., Li, Y., Li, C., Zhang, W., Zhang, Y.E., Lee, U., Zhang, L., Long, M., 2019. Topological evolution of coexpression networks by new gene integration maintains the hierarchical and modular structures in human ancestors. Science China Life Sciences, 62: 594-608.

Journal Pre-proof

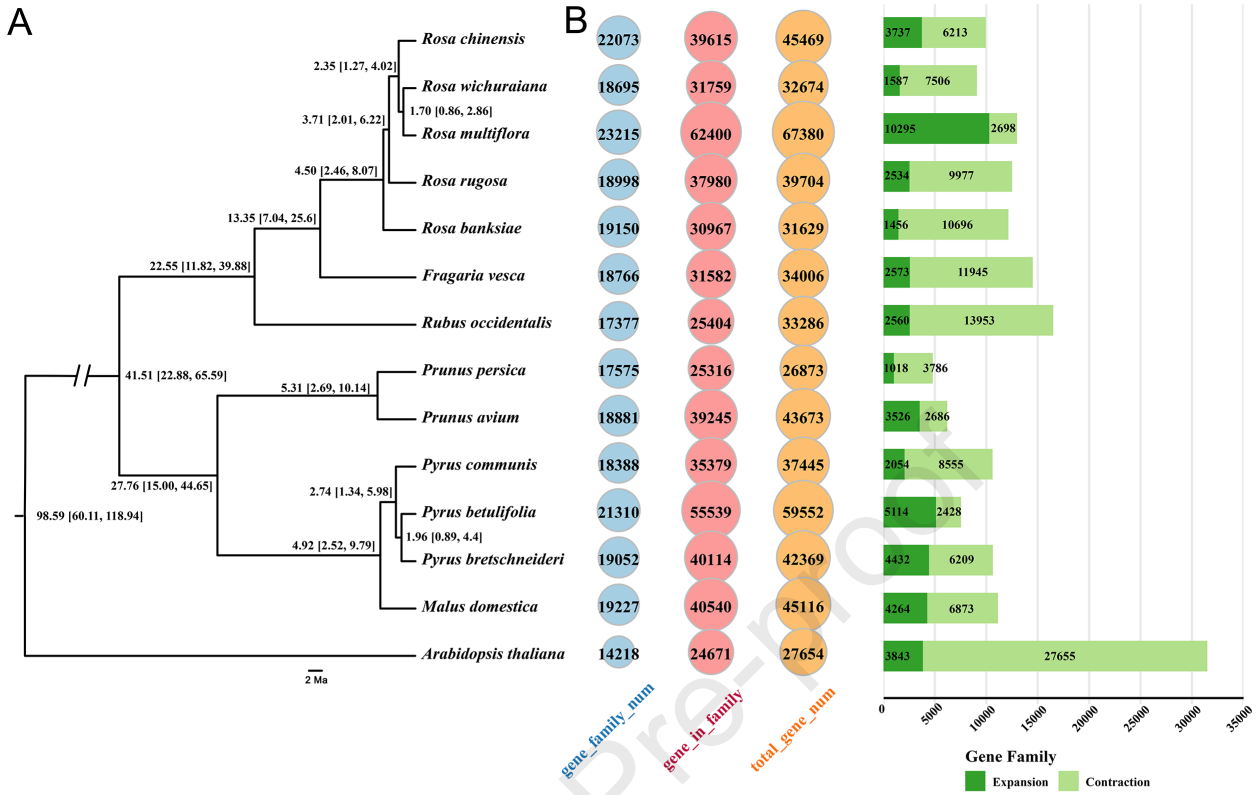


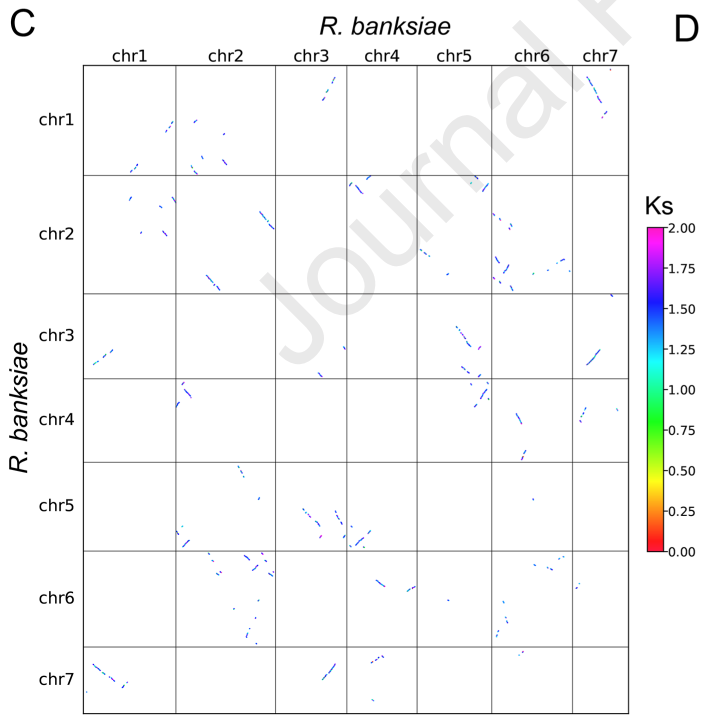
Journal Pre-proof

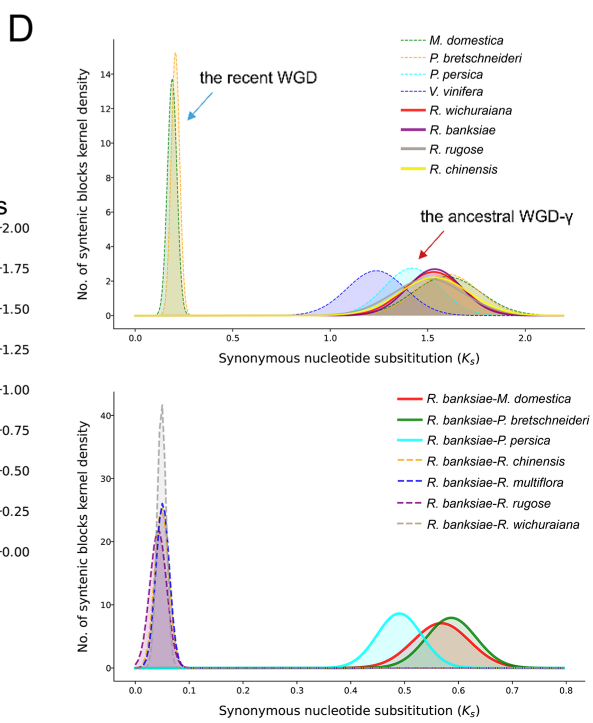


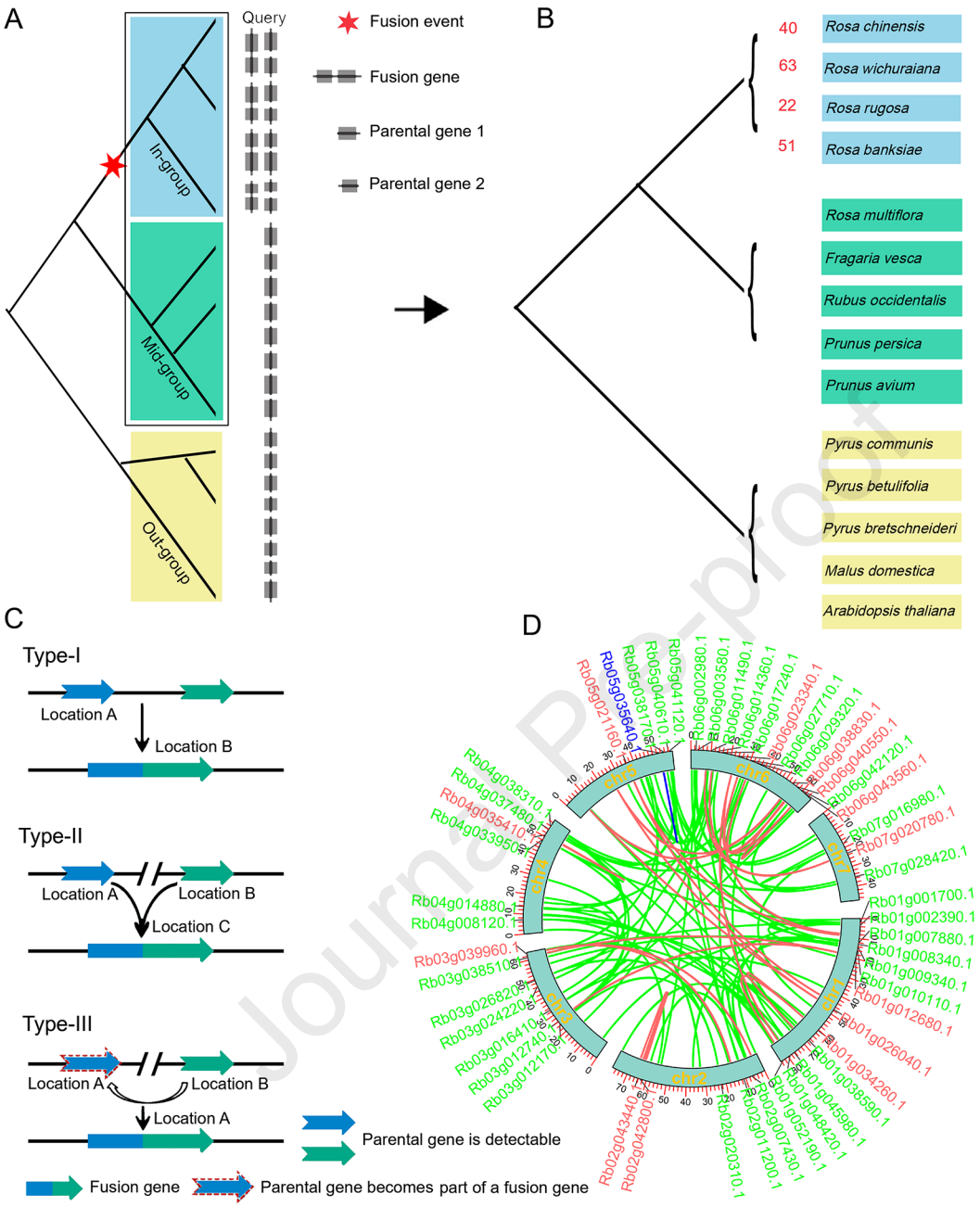
Journal Pre-proof

Journal Pre-proof

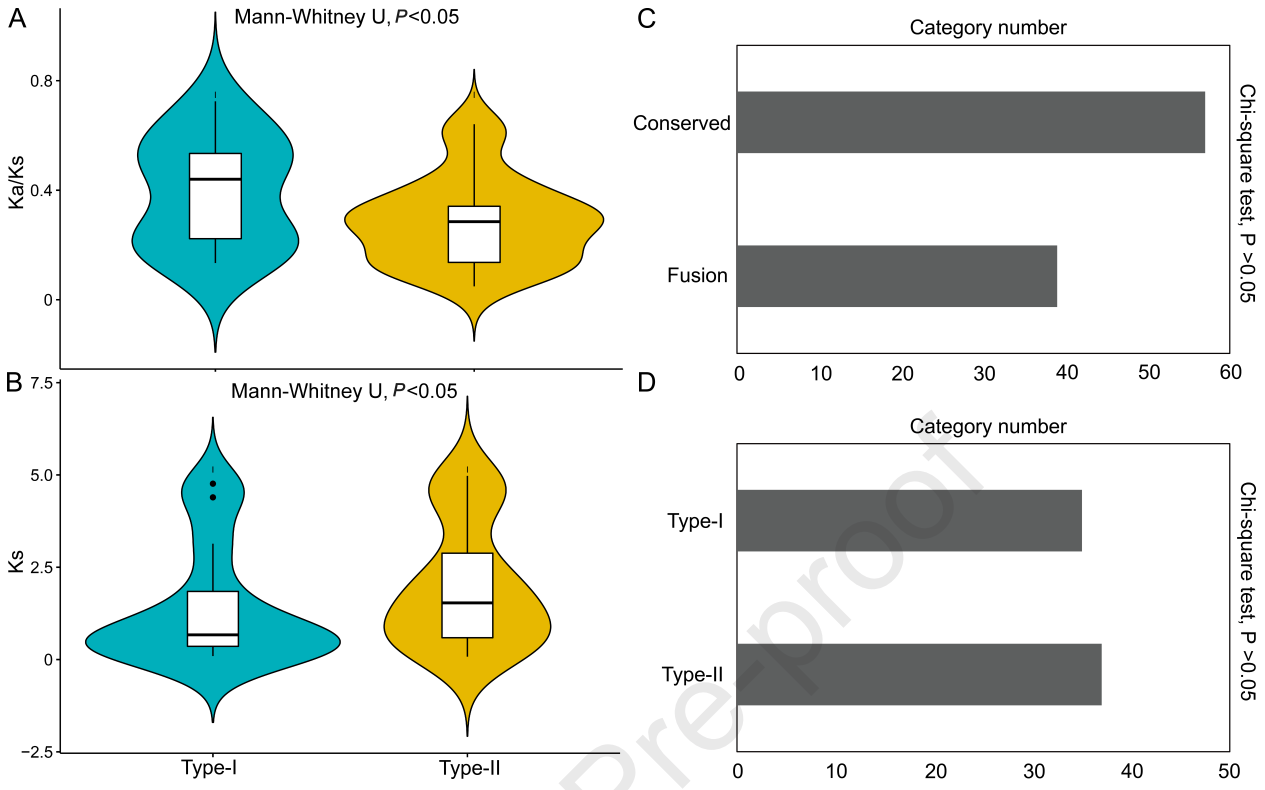




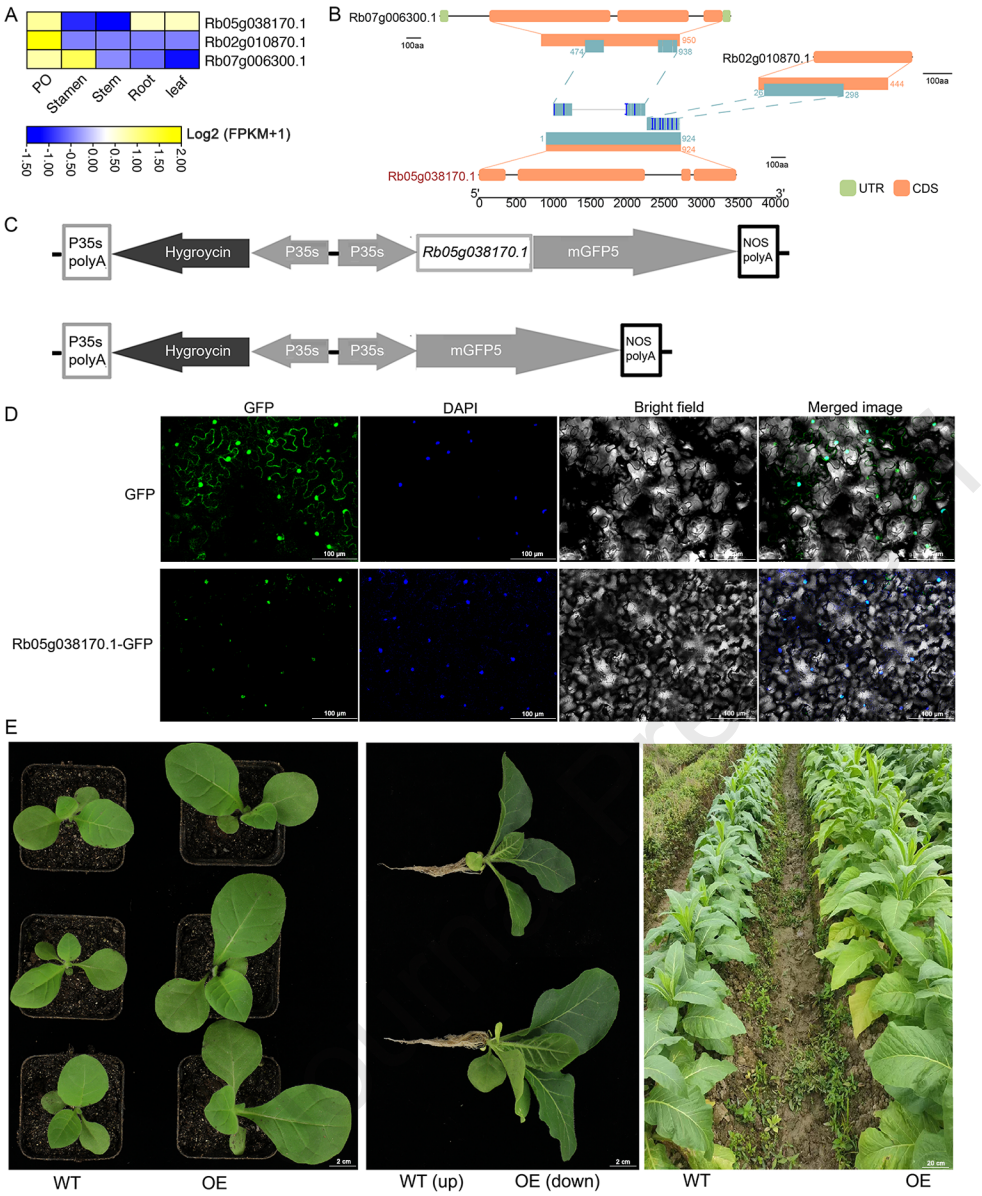




Journal Pre-proof



Journal Pre-proof


A
Rb05g038170.1
Rb02g010870.1
Rb07g006300.1
PO
Stamen
Stem
Root
leaf
Log2 (FPKM+1)
-1.50
-1.00
-0.50
0.00
0.50
1.00
1.50
2.00
B
Rb07g006300.1
100aa
950
474
938
Rb02g010870.1
444
25
298
100aa
1
924
924
Rb05g038170.1
100aa
5'
0
500
1000
1500
2000
2500
3000
3500
4000
3'
UTR
CDS
C
P35s polyA
Hygroycin
P35s
P35s
Rb05g038170.1
mGFP5
NOS polyA
P35s polyA
Hygroycin
P35s
P35s
mGFP5
NOS polyA
D
GFP
DAPI
Bright field
Merged image
GFP
Rb05g038170.1-GFP
100 μm
E
WT
OE
WT (up)
OE (down)
WT
OE
2 cm
2 cm
20 cm

Journal Pre-proof

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Journal Pre-proof