Chromosome-level genome assembly of a parent species of widely cultivated azaleas

Fu-Sheng Yang1,12, Shuai Nie2,12, Hui Liu2, Tian-Le Shi2, Xue-Chan Tian2, Shan-Shan Zhou2, Yu-Tao Bao2, Kai-Hua Jia2, Jing-Fang Guo2, Wei Zhao2,3, Na An1,4, Ren-Gang Zhang5, Quan-Zheng Yun5, Xin-Zhu Wang5, Chanaka Mannapperuma6, Ilga Porth7, Yousry Aly El-Kassaby8, Nathaniel Robert Street6, Xiao-Ru Wang2,3, Yves Van de Peer9,10,11 & Jian-Feng Mao2,3

Azaleas (Ericaceae) comprise one of the most diverse ornamental plants, renowned for their cultural and economic importance. We present a chromosome-scale genome assembly for *Rhododendron simsii*, the primary ancestor of azalea cultivars. Genome analyses unveil the remnants of an ancient whole-genome duplication preceding the radiation of most Ericaceae, likely contributing to the genomic architecture of flowering time. Small-scale gene duplications contribute to the expansion of gene families involved in azalea pigment biosynthesis. We reconstruct entire metabolic pathways for anthocyanins and carotenoids and their potential regulatory networks by detailed analysis of time-ordered gene co-expression networks. MYB, bHLH, and WD40 transcription factors may collectively regulate anthocyanin accumulation in *R. simsii*, particularly at the initial stages of flower coloration, and with WRKY transcription factors controlling progressive flower coloring at later stages. This work provides a cornerstone for understanding the underlying genetics governing flower timing and coloration and could accelerate selective breeding in azalea.

1State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, 100093 Beijing, China. 2Beijing Advanced Innovation Center for Tree Breeding by Molecular Design, National Engineering Laboratory for Tree Breeding, College of Biological Sciences and Technology, Beijing Forestry University, 100083 Beijing, China. 3Department of Ecology and Environmental Science, Umeå Plant Science Centre, Umeå University, SE-901 87 Umeå, Sweden. 4University of Chinese Academy of Sciences, 100049 Beijing, China. 5Beijing Ori-Gene Science and Technology Co. Ltd, 102206 Beijing, China. 6Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, SE-901 87 Umeå, Sweden. 7Département des Sciences du Bois et de la Forêt, Faculté de Foresterie, de Géographie et Géomatique, Université Laval Québec, Québec, QC G1V 0A6, Canada. 8Department of Forest and Conservation Sciences, Faculty of Forestry, University of British Columbia, Vancouver, BC V6T 1Z4, Canada. 9Department of Plant Biotechnology and Bioinformatics and VIB Center for Plant Systems Biology, Ghent University, 9052 Ghent, Belgium. 10Centre for Microbial Ecology and Genomics, Department of Biochemistry, Genetics and Microbiology Genetics, University of Pretoria, Private bag X20, Pretoria 0028, South Africa. 11College of Horticulture, Nanjing Agricultural University, 210095 Nanjing, China. 12These author contributed equally: Fu-Sheng Yang, Shuai Nie. *Email: yves.vandepeer@psb.ugent.be; jianfeng.mao@bjfu.edu.cn*
The genus *Rhododendron* (Ericaceae) harbors more than 1000 species and 30,000 cultivars and is known for the outstanding beauty and diversity of its corollas\(^1\)\(^2\). During the 18th century, several Chinese azalea cultivars were introduced to Europe (especially to England and Belgium), facilitating their breeding for ornamental use\(^3\). *R. simsii* (potted azalea) is an evergreen species of the subgenus *Tsutsusi*, endemic to East Asia, and is the most widely cultivated *Rhododendron* species with numerous cultivars selected from a diversity of wild relatives and natural hybrids\(^5\). At present, the annual production of *R. simsii* hybrids has reached approximately 40 million pots in Belgium alone\(^4\)\(^6\). Although *R. simsii* natural blooming extends from the end of spring to the beginning of summer, some 'easy-care and color-rich' cultivars can bloom as early as Christmas and Spring Festivals, when few other flowering plants are available\(^2\)\(^3\). Currently, azalea cultivars have become the focus of intensive ornamental use through hybridizations and production programs, attaining prominence as pot plants and landscape shrubs in Europe, North America, and Asia\(^3\).

Breeding and development of new ornamental cultivars is mainly focused on flower color. Moreover, flower color is also of paramount importance for plant ecology and evolution\(^3\). Since the work of Gregor Mendel (1856–1863), flower color has successfully contributed to the elucidation of the fundamental principles of genetics, while more recently, genomics has provided insights into the evolution of the biochemical pathways and regulatory networks underlying phenotypic traits, including flower color\(^5\). So far, pigment analyses for azalea flowers were based on pigment type and composition percentage; however, their underlying genetic and regulatory mechanisms are largely unknown. Recently, some of the flavonoid biosynthesis structural genes were isolated from several *Rhododendron* species and cultivars, and the spatiotemporal expression patterns of some key node-genes were analyzed\(^3\)\(^4\)\(^6\)\(^8\). Although two *Rhododendron* genomes were released recently, both were obtained by Illumina short read sequencing\(^7\)\(^8\). The lack of a high-quality *Rhododendron* species whole-genome sequence has seriously hampered the unraveling of their color formation, in spite of the known long-term breeding history\(^3\)\(^8\).

Here, we present a chromosome-scale genome assembly for *R. simsii*. The genome of *R. simsii* is determined by a combination of long-read sequencing and Hi-C scaffolding technologies. In total, a 529 Mb genome sequence is assembled, and >91% of the genome could be anchored on 13 chromosomes with a scaffold N50 of 36 Mb. We detect the remnants of a whole genome duplication and find tandem and other small-scale gene duplications to be the key drivers for gene family expansions. Furthermore, we unravel the metabolic co-expression network of flower pigmentation and identified the structural genes, and their potential regulators, of flower coloring through time-ordered comparative transcriptome analyses. The availability of this reference genome sequence and information on the molecular basis and the genetic mechanisms governing flower color in *Rhododendron* present valuable resources for the development of consumer-oriented selective breeding novelties of azalea.

**Results**

**R. simsii genome assembly and annotation.** The genome size of *R. simsii* (Supplementary Fig. 1) estimated with K-mer analysis to be ~525 Mb (Supplementary Fig. 2) was larger than its size estimation by flow cytometry for yet unresolved reasons (Supplementary Table 1). We produced 100× coverage of PacBio long-read sequencing data, 170× coverage of short reads of PCR-free Illumina sequencing, and 100× coverage of Hi-C paired-end reads (Table 1 and Supplementary Fig. 3). After primary assembly, comparison, correction, polishing, and scaffolding, a final assembly of 529 Mb was obtained. The assembly is slightly larger than the estimated genome size, which may be due to high heterozygosity (~1.78%, estimated with K-mer frequency, see Supplementary Note 1 for details). After mapping the Illumina reads to the final assembly, single nucleotide polymorphisms (SNPs) were identified with SAMtools\(^8\) (with default settings) and obtained a SNP heterozygosity level of ~1.07% and a single base error rate of ~0.0054% was obtained. There was no obvious GC bias in the sequencing data from PacBio single-molecule real-time (SMRT) technology; however, a GC bias was detected for the Illumina sequencing data (Supplementary Fig. 4), confirming the advantage of PacBio SMRT technology over Illumina for genome sequencing\(^10\), in addition to the longer read length provided.

The final assembly consisted of 911 contigs and 552 scaffolds (13 chromosome-level scaffolds, one chloroplast genome, one mitochondrial genome, and 537 super contigs) with contig N50 of 2.2 Mb and scaffold N50 of 36 Mb (Table 1, Supplementary Tables 2, 3 and Supplementary Fig. 5). The number of chromosome-scale super scaffolds is consistent with the species' determined chromosome number of 13\(^11\). The high fidelity of the assembly was supported by the high 10-fold minimum genome coverage of 99.3% (Illumina) and 98.6% (PacBio), and the high mapping rates of 93.3% (Illumina) and 90.9% (PacBio). The high completeness of this assembly was also evidenced by a 93.7% (1349 genes) Benchmarking Universal Single Copy Orthologs (BUSCO) recovery score\(^12\) (Table 1 and Supplementary Table 4), which is better compared to two recently released *Rhododendron* genomes\(^7\)\(^8\), obtained by Illumina short read sequencing (Supplementary Fig. 6). Judged by the high long terminal repeat (LTR) Assembly Index (LAI)\(^13\) score of 18.10, the *R. simsii* genome attained reference level quality.

We predicted 34,170 genes, and the average lengths for total gene regions, transcript, coding sequence (CDS), exon sequence, and intron sequence are 5089.2, 1416.3, 1288.7, 259.7, and 403.1 bp, respectively (Table 1 and Supplementary Table 5). A total of 32,999 protein-coding genes were predicted, which is considerably more than those annotated for the *R. williamsianum* genome (23,559 genes)\(^14\), but similar to those for the *R. delavayi* genome (32,938)\(^15\) (Table 1). In addition, we predicted 482 tRNAs, 64 rRNAs including eight 28S, six 18S and 50 5S rRNAs, and 625 other non-coding RNAs (211 miRNAs, 16 tRNAs and 158 snoRNAs) (Table 1 and Supplementary Table 6).

Among the predicted protein-coding genes, 96.44% could be annotated through at least one of the following protein-related databases: the NCBI non-redundant protein database (NR) (85.70%), the Swiss-Prot protein database (57.80%), the Translated European Molecular Biology Laboratory (TrEMBL) database (84.90%), the protein families database (Pfam) (73.60%), and the Gene Ontology (GO) database (57.54%) (Table 1 and Supplementary Table 7).

**R. simsii is an ancient polyploid.** A phylogenetic tree was constructed for the 15 species of asterids and two outgroup species, using 806 orthogroups (see “Methods” section). The phylogenetic relationship between and within the main clades (Asterid I, Asterid II, and Ericales) agree with previous studies\(^14\)\(^–\)\(^17\). Molecular dating suggests that *R. simsii* diverged from the most recent common ancestor of *R. delavayi* and *R. williamsianum* around 14.54 Ma, following the divergence of *Rhododendron* and *Vaccinium corymbosum*\(^18\) around 55.93 Mya (Fig. 1a).

Synonymous substitutions per site (Ks) age distributions and collinearity/synteny analyses unveiled evidence for an ancient whole-genome duplication (WGD) event in the lineage leading to *Rhododendron*, around 78 Mya (Fig. 1a) (see “Methods” section)
for details), with three Rhododendron species (R. simsii, R. delavayi, and R. williamsianum) showing signature Ks peaks at about 0.65 (Fig. 1b lower left). Moreover, intra-genomic analysis showed that 18.83% of the genome lies within duplicated segments (6,213/32,999 collinear genes distributed along 289 duplicated blocks; Supplementary Table 8). We also built a gene relative rate test, we found that species from Ericales and the number of substitutions per synonymous site per year (Ks) showing signature peak predicted at ~2 Mya for most clades of both Gypsy and Copia (Supplementary Fig. 9 and Supplementary Table 10), which may be one reason for the overall lower proportion of TE in R. simsii than in other species analyzed here.

A total of 825 intact Gypsy (8.68 Mb) and 1303 intact Copia (8.20 Mb) LTR-TRs were identified across the whole R. simsii genome, with recent bursts, as well as a single ancient amplification peak predicted at ~2 Mya for most clades of both Gypsy and Copia (Supplementary Fig. 9 and Supplementary Table 10). Overall, we found 10.5-84.1% among the different subgroups of Gypsy and Copia preferentially residing in gene regions, particularly the chloroplast RNA splicing and ribosome

### Table 1 The statistics for genome sequencing of Rhododendron.

|                        | R. simsii | R. delavayi | R. williamsianum |
|------------------------|-----------|-------------|------------------|
| **Sequencing**         |           |             |                  |
| Raw bases of WGS-PacBio Seq (Gb) | 51.15     | *           | *                |
| Raw bases of WGS-Illumina (Gb)     | 91.49     | 336.83     | *                |
| Raw bases of Hi-C (Gb)       | 55.681    | *           | *                |
| Raw bases of mRNAseq (Gb)    | 422.149   | *           | *                |
| **Assembly**             |           |             |                  |
| Genome size (Mb)          | 528.6     | 695.1       | 532.1            |
| Number of scaffolds       | 552       | 193,091     | 11,985           |
| N50 of scaffolds (bp)     | 36,350,743| 637,826     | 218,828          |
| L50 of scaffolds          | 7         | *           | *                |
| Chromosome-scale scaffolds (bp) | 481,946,564 (91.17%) | 0 | 368,385,547 (69%) |
| Number of contigs         | 911       | *           | *                |
| N50 of contigs (bp)       | 2,234,511 | 61,801      | *                |
| L50 of contigs            | 66        | *           | *                |
| Number of Gap             | 359       | *           | *                |
| Complete BUSCOs           | 93.68%    | 92.80%      | 89%              |
| GC content of the genome (%) | 38.91%   | *           | *                |
| **Annotation**            |           |             |                  |
| Number of predicted genes | 34,170    | *           | *                |
| Number of predicted protein-coding genes | 32,999 | 32,938 | 23,559 |
| Average gene length (bp)  | 5,089.22  | 4,434.22    | 4,628            |
| Average CDS length (bp)   | 1,288.73  | 1,153.21    | *                |
| Average exon per transcript | 5         | 4.62        | 5.68             |
| Number of rRNAs           | 482       | *           | *                |
| Number of rRNAs           | 64        | *           | *                |
| Repeat sequences (bp)     | 250,988,768 (47.48%) | 359,874,503 (51.77%) | * |
| Annotated to Swiss-Prot   | 19,079    | 22,693      | *                |
| Annotated to PFAM         | 24,301    | *           | *                |
| Annotated to GO           | 25,038    | 16,471      | 18,538 (79%)     |
| Annotated to KO           | 11,506    | (34.90%)    | *                |

The assembly for R. simsii was compared with two previously reported genome assemblies for Rhododendron species, R. williamsianum, and R. delavayi. Asterisk (*) indicates data were not shown in the original articles.
Fig. 1 Genome evolutionary history. 

a Chronogram showing divergence times and genome duplications in asterids (asterid II, asterid I, Ericales, and Cornales), with node age and the 95% confidence intervals labeled. Resolved polyploidization events are shown with blue (duplications) and red translucent dots (triplications). Pie charts show the proportions of gene families that underwent expansion or contraction. Predicated WGD only shown for Ericales and Cornales. *WGD identified in this study. †WGD reported in Wang et al.59. ‡WGD reported in Colle et al.18. b Ks distribution on the upper right (insert) is showing Ks distribution from orthologs between Vitis vinifera and each of the five species (Rhododendron delavayi, R. simsii, R. williamsianum, Camellia sinensis, and Camptotheca acuminata). Ks distribution in the lower left showing Ks distribution from paralogs within Rhododendron delavayi, R. simsii, R. williamsianum, Camellia sinensis, and Camptotheca acuminata. c Synteny and distribution of genomic features. a, the density of Ogre, a family of Gypsy LTR-RT. b, the density of Tekay, a family of Gypsy LTR-RT. c, The density of all Gypsy LTR-RT. d, the density of all Copia LTR-RT. e, gene density. f, histogram of GC content. A map connecting homologous regions of the genome is shown. The red lines represent syntenic regions for the WGD event (78 Mya) and blue lines represent γ-WGD event. The line segments of bold gray in outer circles indicate predicted centromeres and pericentromeric regions on the 13 chromosomes.
maturation (CRM) family of Gypsy (57.63%) and the Ale family of Copia (52.73%), Supplementary Figs. 10a, b and Supplementary Table 11. However, most members of Gypsy and Copia were found 3-5 Kb distant from genic regions (Supplementary Fig. 10c). Insertion dynamics were similar between LTR-TRs proximal to and far away from genes, and among subgroups of both Gypsy and Copia (Supplementary Figs. 10d, e and l).

Gene density decreased and GC content increased from the chromosome ends towards the centromeres (Fig. 1c and Supplementary Fig. 11). We found that the accumulated Gypsy elements tend to be clustered in the pericentromeric regions, particularly Tekay and Ogre, yet there was a relatively even distribution for Copia elements along the chromosomes. Moreover, it was found that more Copia elements experienced positive selection (Ka/Ks > 1) in their reverse transcriptase (RT) domains, while RT domains of Gypsy elements showed lower Ka/Ks ratios and relatively smaller Ka values (Supplementary Fig. 12). The distinct behaviors of Gypsy and Copia remain to be elucidated.

**TD/PD contributed to gene family expansion in R. simsii.** We identified 29,396 duplicated genes that were classified into five different categories: 6056 whole-genome duplicates (WGD duplicates, 18.4%), 4746 tandem duplicates (TD, 14.4%), 3732 proximal duplicates (PD, 11.3%), 6399 transposed duplicates (TRD, 19.4%), and 8463 dispersed duplicates (DSD, 25.6%) (see “Methods” section, Supplementary Fig. 13). We compared the Ka and Ks distribution among different modes of gene duplication. Higher Ka/Ks ratios and smaller Ks values were found for tandem and proximal duplicate gene pairs (Fig. 2a and Supplementary Fig. 14), suggesting an ongoing and continuous process for tandem and proximal duplications and more rapid sequence divergence and stronger positive selection than genes originated through other duplication modes.

All predicted gene models for the 17 species were clustered using OrthoFinder (version 2.3.1)\(^{22}\), resulting in 22,455 orthogroups (Supplementary Tables 12 and 13). Using CAFÉ (version 4.0)\(^{23}\), 1515 gene families (6,754 genes) were found to be expanded, while 16575 gene families were found to be contracted in R. simsii, with 57.6% of the expanded gene families (EGFs) due to TD and PD duplications (Fig. 2b). Hypergeometric tests were performed for overlapping genes between expanded gene families and WGD-TD-PD-TR-DSD (Fig. 2c). TD-EGFs and PD-EGFs were found exhibiting divergent enriched GO terms. For example, TD-EGFs enriched categories are implicated in plant self-defense, development and adaptation, while PD-EGFs genes are enriched in GO terms involved in ‘binding’ and ‘recognition’. In brief, newly generated tandem and proximal duplications have been important sources of gene family expansion in R. simsii (Supplementary Fig. 14).

To verify whether the identified tandem gene clusters are real and not artificial due to errors in the genome assembly, we mapped our PacBio long-reads back to the assembly, and examined whether the pair of adjacent tandem duplicated genes or the intergenic region could be fully or partially recovered by the mapped long-reads. We indeed found that most (79–87%) of the duplicated genes could be recovered, fully or partially, both for transcription factor (TF) or anthocyanin/flavonol biosynthetic genes (Supplementary Fig. 15). These results provide very good evidence for the true existence of tandem gene clusters.

**Flowering-time genes in R. simsii.** We detected 424 genes related to flowering-time control in R. simsii by querying the Flowering Interactive Database, FLOR-ID\(^{24}\), in which 295 flowering-time genes are functionally characterized for Arabidopsis. With regards to gene function, the categories ‘General’ and ‘Photoperiodism, light perception and signaling’ were represented in large proportions of 37.26% (158) and 20.99% (89) among all flowering-time control genes (Fig. 2d). There was a clear time-gradient of expression of the flowering-time related genes (Supplementary Fig. 16) across the five examined flower coloring stages. In addition, we identified a gene family (OG0000614) encoding 13 members of high-affinity sucrose transporters that may play key roles in flowering transition delay\(^{25}\). This gene family was identified by CAFÉ as being expanded in azalea, and its members showed continuous upregulation in the corolla during the flowering time-series (Supplementary Fig. 16).

**Flower pigmentation genes in R. simsii.** Our high-quality genome assembly allowed reconstruction of the metabolic pathway for flower coloration, specifically capturing the implicated enzymatic genes in this process. We unveiled 58 genes encoding enzymes functioning in the carotenoid biosynthesis pathway (Supplementary Fig. 17) and 125 enzymatic genes with predicted functions in anthocyanin and flavonol biosynthesis (Fig. 3 and Supplementary Fig. 18). Flavonoids are synthesized by a branched pathway that yields both colored anthocyanin pigments and colorless flavonols. Furthermore, we predicted five genes encoding three enzymes of blue-anthocyanins modification, including one anthocyanin 5-(6’-hydroxy)cinnamoyltransferase (5AT|3AT, one of gentiodelphin-acyltransferase, EC 2.3.1.153), two ternatin C3-acyltransferases (ternatin C3-AT, EC 3.2.2.24), and two viodelphin-glucosyltransferases (viodelphin-GT, EC 2.3.1.2.-) (Fig. 3 and Supplementary Fig. 18).

**Gene duplication events linked to flowering time and color.** TD genes were found to be enriched for GO terms related to secondary metabolic biosynthetic and modification processes (Fig. 2c). Furthermore, the proportion of TD/PD genes linked to flower coloration was significantly higher in R. simsii than that in related species (Supplementary Table 14). Counting the number of enzymatic genes in tandem or proximal gene clusters unveiled that TD/PD duplications substantially contributed to the observed proportions of enzymatic genes for the anthocyanin/flavonol versus the carotenoid biosynthesis pathway. TD/PD duplications affected 17% of genes for carotenoid biosynthesis in the R. simsii genome (Supplementary Fig. 17 and Supplementary Table 15) and 45% of the genes for anthocyanin/flavonol biosynthesis (Fig. 3 and Supplementary Table 15), respectively. In addition, important TD/PD events were also found for 52.63% (10/19) of the genes coding for 4CL (4-coumarate: CoA ligase 4; EC 6.2.1.12), 57.14% (8/14) encoding CHS (chalcone synthase; EC 2.3.1.74), 60% (3/5) encoding DFR (dihydroflavonol 4-reductase; EC 1.1.1.219), 48.48% (16/33) encoding F3H (flavonone-3-hydroxylase; EC 1.14.11.9), 35.29% (6/17) encoding FLS (Flavonol synthase; EC 1.14.11.23), 55.56% (10/18) encoding F3′H (flavonoid 3′-hydroxylase; EC 1.14.13.21), 58.33% (7/12) encoding F3′5′H (flavonoid 3′,5′-hydroxylase; EC 1.14.13.88) (Fig. 3). The diversity of the expression patterns indicates that further studies are necessary to verify and characterize the functions of the TD/PD duplicates.

In contrast, TD and PD duplications seem to have accounted for much lower proportions of expansions of flowering-time genes (6.82% (30/424) and 6.84% (424), respectively)\(^{24}\), respectively), while WGD appears to have contributed the largest proportion (159/424, 37.5%) of flowering-time control gene copies and more than other modes of gene duplication (Fig. 2d). Interestingly, genes involved in flowering-time control are mostly present as low-copy gene families, 101 genes from single-copy families, 128 genes from dual-copy families and 60 genes from triple-copy families (Supplementary Fig. 19).
Fig. 2 Gene duplication and evolution. a The Ka/Ks ratio distributions of gene pairs derived from different modes of duplication. WGD whole-genome duplication, TD tandem duplication, PD proximal duplication, TRD transposed duplication, DSD dispersed duplication. In the boxplot representations, points are outliers; the center line is the median; the lower and upper hinges correspond to the first and third quartiles (25th and 75th percentiles); whiskers extend to the minimum (left whiskers) and maximum (right whiskers) estimates located within the 1.5× interquartile range (IQR) of the first and third quartiles, respectively. Gaussian kernel estimates of Ka/Ks of different duplicated groups are shown as violins; 6056 WGD, 4746 TD, 3732 PD, 6399 TRD, and 8463 DSD duplicates were incorporated in the statistics. b Venn diagram shows the possible logical relations between members of expanded gene family and modes of duplication (WGD, TD, PD, TRD, and DSD). TDp: tandem duplication experienced positive selection, PDp: proximal duplication experienced positive selection, DSDp: dispersed duplication experienced positive selection, EGFs: expanded gene families. c Functional enrichment analysis of genes overlapping between expanded gene families (EGFs) and five modes of duplication (WGD, TD, PD, TRD, and DSD). The enriched GO terms with corrected P value <0.001 are presented. The color of circles represents the statistical significance of enriched GO terms. The size of the circles represents the number of genes in a GO term. For all annotated genes their GO terms are provided as background information. ‘P adjust’ is the Benjamini–Hochberg false discovery rate (FDR) adjusted P value. d The numbers of genes from different families and gene duplications are provided. SL singletons.
The dynamics of flower pigmentation regulation in *R. simsii*. We performed time-ordered comparative transcriptome analyses (Fig. 4a and Supplementary Fig. 20) and found 8067 genes (618 TFs and 7,449 structural genes) expressed with an average Transcripts Per Kilobase Million (TPM) greater than 0.5 and exhibiting significant differentiation between any two samples among the five flowering time stages (T1–T5). As the initial node, a basic helix-loop-helix (bHLH) transcription factor (*Rhsim13G0024200*), highly expressed at the very first time point but not later, was selected to generate a time-ordered gene co-expression network (TO-GCN)⁴⁶. Eight time-series expression levels (L1–L8, nodes >20) centering on TFs were finally reconstructed using the suggested positive/negative cutoff values (0.81; –0.57) (Fig. 4b). TO-GCN revealed a co-expression network involving the 618 differentially expressed TFs and 62 genes from the carotenoid/anthocyanin/flavonol biosynthesis pathways (Fig. 4b). Time-course transcription factors were further distinguished for the initial (T1, flower appears white and semi-transparent), the transitional (T2 and T3 for light red flower), and the terminal (T4 and T5 for the determined red flower color) stages of corolla pigmentation (Fig. 4c).

The general pattern elucidated that most network members (358, including 328 TFs) appeared at the initial stage (T1), 177 genes (including 150 TFs) at the terminal stage (T4–T5), and 70 genes (65 TFs) at the transitional stage (T2–T3). Likewise, T1 (corresponding to time-series levels L1–L3) showed the highest TPM, followed by genes assigned to the transitional stage (T2–T3, corresponding to L4–L5), and finally the genes that were expressed at the terminal red color determining stage (T4–T5, corresponding to L6–L8). Functional enrichment analysis revealed that distinct gene functions are turned on and off, respectively, at these different stages (Supplementary Fig. 21). Yet, some genes encoding key biosynthetic enzymes in color determination tend to be continuously upregulated, such as those in the carotenoid biosynthesis pathway encoding phytoene synthase (*PSY*; EC 2.5.1.32), carotene isomerase (*CRTISO*; EC 5.5.1.6), F3H flavanone 3′-hydroxylase (EC 1.14.14.19), F3′5′H flavonoid 3′,5′-hydroxylase (EC 1.14.13.88), DFR dihydroflavonol 4-reductase (EC 1.1.219), ANS anthocyanidin synthase (EC 1.14.11.19), F3′5′H flavonol 3′,5′-hydroxylase (EC 1.4.13.88), AT acyltransferase, GT glucosyltransferase, FLS flavonol synthase (EC 1.14.11.23). Gene expression profile (in normalized TPMs) in different time points of flowering (here T1-T5, from left to right in each heatmap panel) are presented in the heatmap alongside the gene names. The bar represents the expression level of each gene (z-score). Low to high expression is indicated by a change in color from blue to red.

**Fig. 3** The biosynthesis pathways of two kinds of flavonoids, anthocyanin and flavonol. Genes identified in tandem clusters are displayed as blue while proximal cluster genes are displayed as brown. We divided the flavonoid enzymes encoding genes into two groups designated as ‘early’ and ‘late’, respectively⁷⁰. The ‘early’ biosynthesis genes (EBGs) encode enzymes functioning more upstream towards the entrance into the pathway, whereas the ‘late’ biosynthesis genes (LBGs) exhibit highly distinct expression patterns. PAL phenylalanine ammonia-lyase (EC 4.3.1.24), C4H: cinnamate-4′-hydroxylase (EC 1.14.13.21), F3H flavanone 3′-hydroxylase (EC 1.14.19), F3′5′H flavonoid 3′,5′-hydroxylase (EC 1.14.13.21), DFR dihydroflavonol 4-reductase (EC 1.1.219), ANS anthocyanidin synthase (EC 1.14.11.19), F3′5′H flavonol 3′,5′-hydroxylase (EC 1.14.13.88), AT acyltransferase, GT glucosyltransferase, FLS flavonol synthase (EC 1.14.11.23). Gene expression profile (in normalized TPMs) in different time points of flowering (here T1-T5, from left to right in each heatmap panel) are presented in the heatmap alongside the gene names. The bar represents the expression level of each gene (z-score). Low to high expression is indicated by a change in color from blue to red.
levels (Fig. 5a and Supplementary Table 16). Among the 194 TFs that appeared as potential direct regulators of these enzymatic genes, WD40 family members were most numerous (21). For anthocyanin/flavonol biosynthesis, we detected 16 enzymatic genes expressed at high levels at T1 (Fig. 5b and Supplementary Table 16), and 120 TFs as their potential direct regulators. For genes implicated in the sub-network (Fig. 5c) for anthocyanin/flavonol biosynthesis and their associated bHLH, WD40, and MYB TFs (Supplementary Table 17), a clear gene expression decline was evident (Supplementary Fig. 22), exemplified for some of these TFs and their direct target, a dihydrokaempferol biosynthetic gene F3H (Rhsim11G0126300) (Fig. 5b–d).

The terminal coloration stage (T4–T5) showed maximal expression for substantial numbers of TFs and pigment biosynthetic genes. For anthocyanin/flavonol biosynthesis, we predicted that seven enzymatic genes (Supplementary Table 16 and Fig. 6a) expressed at high levels were directly regulated by 74 potential regulators, mostly ethylene-responsive element binding factors (ERFs) (19) and WRKY (16) family members. The phylogenies, conserved motifs and gene structures of these TFs associated with flower coloration were further examined (Supplementary Figs. 23 and 24).

F3H represents an important rate-limited enzyme in anthocyanins/flavonols synthesis, but its upstream regulators have remained elusive. By examining the co-expression network inferred from TO-GCN (Fig. 6b, c), we identified the potential first-order to third-order upstream regulators. Here, we identified one F3H gene (Rhsim03G0111400) as the hub within the potential regulatory network relationships involving 14 TFs (Fig. 6c). We could infer that this F3H gene may be regulated in a hierarchical order by WRKY Rhsim05G0226500 as the third regulator, either GRAS Rhsim13G0080100 or C3H Rhsim13G0068400 as the second regulator, and MYB Rhsim08G0132300 or C2H2 Rhsim10G0164300 as the direct regulator. When we incorporated the DNA binding site predictions, we found C2H2 may be the sole intermediary regulator in the hierarchy, since only a C2H2 binding site was identified 5’ upstream of the F3H gene. By comprehensively examining the putative hierarchical gene regulation for all key enzymatic genes in anthocyanin/flavonol biosynthesis pathways (Fig. 6c and Supplementary Fig. 25), WRKY Rhsim05G0226500 was identified as an important upstream regulator in these pathways.

Discussion

We here describe a chromosome-scale genome assembly for a key parental species of the cultivated azalea, Rhododendron simii (Ericales). In the asterids clade of the eudicot plants, representing the parental species of the cultivated azalea, Rhododendron simsii (16) family members. The phylogenies, conserved motifs and gene structures of these TFs associated with flower coloration were further examined (Supplementary Figs. 23 and 24).

F3H represents an important rate-limited enzyme in anthocyanins/flavonols synthesis, but its upstream regulators have remained elusive. By examining the co-expression network inferred from TO-GCN (Fig. 6b, c), we identified the potential first-order to third-order upstream regulators. Here, we identified one F3H gene (Rhsim03G0111400) as the hub within the potential regulatory network relationships involving 14 TFs (Fig. 6c). We could infer that this F3H gene may be regulated in a hierarchical order by WRKY Rhsim05G0226500 as the third regulator, either GRAS Rhsim13G0080100 or C3H Rhsim13G0068400 as the second regulator, and MYB Rhsim08G0132300 or C2H2 Rhsim10G0164300 as the direct regulator. When we incorporated the DNA binding site predictions, we found C2H2 may be the sole intermediary regulator in the hierarchy, since only a C2H2 binding site was identified 5’ upstream of the F3H gene. By comprehensively examining the putative hierarchical gene regulation for all key enzymatic genes in anthocyanin/flavonol biosynthesis pathways (Fig. 6c and Supplementary Fig. 25), WRKY Rhsim05G0226500 was identified as an important upstream regulator in these pathways.

Discussion

We here describe a chromosome-scale genome assembly for a key parental species of the cultivated azalea, Rhododendron simii (Ericales). In the asterids clade of the eudicot plants, representing ~25% of all flowering plant species, we dated a whole-genome duplication in the Ericales lineage around 78 Mya. We believe this WGD event to be shared between Ericaceae and Actinidiaceae, as
Factors underlying pigment formation for individual genes encoding relevant enzymes and gene transcription beyond genetic constraints that are difficult to overcome through conventional breeding. Previous studies have characterized indigo pigments in several Ericaceae, Actinidiaceae, and Theaceae. TD/PD, known to act as a crucial factor in pigment biosynthesis, was found to be highly enriched in certain gene families. We found that TD/PD duplications in azalea have substantially contributed to the proportions of enzymatic genes for the anthocyanin biosynthesis pathway, highly suggestive of the important role of TD/PD events in shaping flower color diversity.

Genetic modification in pigment biosynthetic pathways would provide a powerful method for obtaining novel flower colors beyond genetic constraints that are difficult to overcome through conventional breeding. Previous studies have characterized individual genes encoding relevant enzymes and gene transcription factors underlying pigment formation for floral crops. Nevertheless, despite the studies on pigment composition for Rhododendron dating back to the 1950s, the molecular basis of azalea flower coloration has remained elusive, since so far, only preliminary expression analysis was performed and for a limited number of structural genes.

Based on the chromosome-scale genome sequence of R. simsii presented here, in combination with the functional mapping and time-ordered transcriptome analysis, we first reconstructed the entire metabolic pathways for anthocyanin/flavonol/carotenoids biosynthesis in Rhododendron. The resolved pathways unravel a MYB-bHLH-WD repeat (MBW) complex composed of MYB, bHLH, and WD40 TFs that collectively regulate flavonoids/biosynthesis in flower tissues. Importantly, the time-ordered gene co-expression networks for the terminal stage of flower development show WRKY and ERF TFs at the center of the co-expression network, with all involved WRKYs belonging to Group II. We speculate that members of these two TF families, known to be responsible for biotic/abiotic stress responses, may also be involved in flower pigment biosynthesis, and likely play vital roles in flower color intensity and patterning.

Flower colors of potted azalea are restricted to red, pink, white, reddish-purple, and purple, and providing yellow floral color has been the dream of many azalea breeders. Here, we reconstructed...
the carotenoid biosynthesis pathway known to be responsible for yellow, orange, or red pigments, as well as several genes encoding key enzymes of blue-anthocyanin biosynthesis responsible for blue pigment, none of which has been reported in *Rhododendron*. Moreover, gene expression for the generation of myricetin, kaempferol, and quercetin was detected as representatives of flavonols, which may also be involved in the flower color formation of *Rhododendron*33. These findings imply that potted azalea has the genetic potential to produce yellow and even extremely rare floral colors, such as violadelphin, and the genetic basis for flower coloring in *Rhododendron* far exceeds our cognition and it has great potential as a model plant for flower color research.

This study uncovered that MBW complexes may regulate *F3H* to control the biosynthesis of dihydroxyflavonols and subsequently also control anthocyanin/flavonol metabolism, suggesting that *F3H* should be a key node gene in regulating flower color formation in azalea. Moreover, the gene co-expression networks for metabolic pathways revealed that a specific WRKY transcription factor may play a core role in regulating flower coloration in azalea. We present a co-expression network for flower coloring. Fresh tissues of five core genes and the potential regulatory elements (here TFs).

In conclusion, the reference genome sequence presented in this study provides a key resource for the further development of hypotheses and techniques in plant biotechnology, such as molecular marker-assisted breeding and genome editing, which are necessary to aid *Rhododendron* horticulture research and increase the efficiency for *Rhododendron* breeding.

**Methods**

*Rhododendron simsii* plant material. We sequenced the entire genome of a 20-year-old *Rhododendron simsii* individual originating from Jiangshu, Hubei Province, China, and performed de novo genome assembly (Supplementary Fig. 1). This shrub had been transplanted to the Botanical Garden, Institute of Botany, Chinese Academy of Science, Beijing, China, and from which different plant tissues were sampled: fully expanded leaves were used for genome sequencing library preparation, while flowers, young leaves, and young stems were sampled for RNA sequencing (RNAseq) in support of genome annotation. Tissues were immediately flash frozen and stored at −80 °C for subsequent nucleic acid extractions.

In addition, corolla samples of five replicate plants from five individuals of *R. simsii* of the wild population selected for genome sequencing were collected to unravel the gene regulatory network underpinning flower coloring. Fresh tissues were first stored in RNAlater (Ambion, Life Technologies, Austin, TX, USA), then conserved at −80 °C.

**Library construction and sequencing of *R. simsii***. Total DNA was isolated and extracted from the leaves using the DNeasy Plant Mini Kit (Qiagen, Inc.) and then purified using the Mobio PowerClean Pro DNA Clean-Up Kit (MO BIO Laboratories, Inc.). For PacBio SMRT sequencing, sheared and concentrated DNA was used to construct sequencing libraries with 20-kb DNA inserts and subsequently run on a PacBio RSII platform using P6-C4 chemistry (6 SMRT cells). After data filtering and preprocessing, a total of 6.5 million PacBio long reads were generated, yielding ~50 Giga bases (Gb) (100× coverage) with an average read length of 7705 base pairs (bp).

For Illumina sequencing, 150-bp paired-end (PE) libraries were prepared for sequencing on an Illumina HiSeq X Ten platform. Short reads were processed with fastq (version 0.19.3)37. Finally, we obtained ~91.49 Gb of raw sequencing data (corresponding to roughly 170× the assembled genome).

One Hi-C library prepared with MboI restriction enzyme was sequenced on Illumina HiSeq X Ten to generate ~55.68 Gb of valid data from 150 PE reads. De novo genome assembly of *R. simsii* was performed using SMARTdenovo (see URLs section) from PacBio long reads. Subsequently, we merged the two optimal assemblies, v0.3 (criteria: reasonably sized assembly, fewest contigs) and v0.4 (criterion: highest contig N50) by quickmerge (version 0.2)40, further polished with one round of pilon (see URLs section) using Clean Illumina reads to generate assembly v1.0. Based on Hi-C data and assembly v1.0, primary scaffolds were produced by 3D-DNA (version 180922) (see URLs section). These scaffolds were roughly split by Juicebox (version 1.8) (see URLs section) and another round of scaffolding by 3D-DNA (version 180922). Afterwards, we elaborately optimized the new scaffolds with gap closing by LR Capcloser (version 1.1) (see URLs section) and five rounds of pilon polishing.

BUSCO and LAI were used to assess genome completeness and continuity. Furthermore, we mapped 51.15 Gb PacBio sequencing reads, 91.49 Gb Illumina reads, 403.67 Gb clean RNA-seq sequences, and 55.68 raw Hi-C data to the genome assembly using BWA-MEM (see URLs section), minimap2 (version 2.1.1-7977)41.

---

**Fig. 6 Sub-network for different pigment biosynthesis pathways in the terminal stage.** a Sub-network for anthocyanin and flavonol biosynthesis. b DNA binding sites detected in the 2 kb upstream sequences of five core genes and the potential regulatory elements (here TFs). c Resolved hierarchical regulation for *F3H* gene. The bar represents the expression level of each gene (z-score). Low to high expression is indicated by a change in color from blue to red. Edges between structural genes were not shown.

---
HiSat2 (version 2.1.0) (see “URLs” section) and Juicer (see “URLs” section), respectively.

Additional details regarding genome assembly and quality control are provided in the Supplementary Note 4.

Annotation of repeat elements and genes for R. simsii. Repeat elements were identified by a combination of homology-based and de novo approaches using RepeatModeler (version 1.0.10) (see “URLs” section) and RepeatMasker (version 4.0.7, trimblast 2.2.28) (see “URLs” section). Furthermore, we examined classification, age, abundance, hierarchy and death of LTR-RTs (transposable element-like repeats) using RepeatExplorer (version 2.2.28+.), BLASTn (version 2.2.28+.) and RepeatModeler (version 1.0.10) (see “URLs” section). Gene functions were annotated according to both homology and similarity searches by blast (version 36)46 with 30% identity and 1e-50 E-value cutoff and domain similarity prediction strategies using InterProScan (version 5.27-66.0) (see “URLs” section).

Locations of all contromers within the genome were inferred with Centurion17 based on corrected Hi-C data and the tendency of formed clusters in three-dimensional space. Additional details are available in the Supplementary Note 5.

Inference and analysis of orthogroups for asterids. To retrieve the evolutionary history of the asterids clade, a total of 22,455 orthogroups of azalea and 16 representative plant species (Supplementary Tables 12 and 13) were identified by using OrthoFinder (version 2.3.1)42. Then, OrthoFinder provides a formalized procedure for determining 806 low-copy orthologs with minimum of 76.5% of species having single-copy genes in any orthogroup. The concatenated amino acid sequences alignment was created by MUSCLE (version 3.8.31)50 and trimmed with trimAl (version 1.2) (trimal -gt 0.8 -st 0.001 -cons 60)49. And then a maximum likelihood (ML) phylogenetic tree was constructed using IQ-TREE (version 1.6.7)50 with the selected optimal sequence evolution model (-m JTT) followed by featureCounts63 to generate the count and Transcripts per Kilobase Million (TPM) of genes. Differential gene expression (DGE) analyses between the five stages were performed with DESeq24, with 0.05 as the FDR cut-off and a log2 fold change (FC) cut-off of 1 (Supplementary Fig. 2).

We used a recently developed method for reconstructing time-ordered gene co-expression networks (TO-GCNs)59. Significantly differentially expressed genes with average TPM less than 0.5 were excluded. The Pearson correlations between 6,915 azalea and 7,449 genes less than 0.5 or above 0.81 in C1 + C2 + GCN were visualized in graphs using Cytoscape65.

Finally, 2K sequences in the upstream of six core genes (Rhizosmis03G011400, Rhizosmis08G0132500, Rhizosmis10G0146300, Rhizosmis11G0126300, Rhizosmis13G0068400, and Rhizosmis13G0080100) were extracted and queried against PlantRegMap and PlantCARE95. TF binding sites unscovered by PlantRegMap were illustrated with R package drawProteins8 (Figs. 5d and 6b). Additional details regarding TFs are provided in the Supplementary Note 8.

Gene functional enrichment. Hypergeometric tests were performed to determine whether specific functional categories from GO were significantly over-represented in R. simsii gene sets within the genome. Functional enrichment was tested using the R package clusterProfiler (version 3.6.0)48 with the background set of all annotated R. simsii genes in this study.

URLs SMARTdemonovo [https://github.com/ruanjie/smartdemonovo]; FALCON-Phase (version 0.1.0-beta) [https://github.com/WGLab/EnhancedFALCON]; Pilon [https://github.com/broadinstitute/pilon]; 3D-NDA (version 180922) [https://github.com/theaidenlab/3d-dna]; Juicebox (version 1.8) [https://github.com/ab荒/juicebox]; LR_Gapcloser [https://github.com/ib3/bwa]; HiSat2 (version 2.1.0) [https://github.com/infphilo/hiSAT2]; Juicer [https://github.com/aidenlab/juicer]; RepeatMasker [http://www.repeatmasker.org]; RepeatModeler [http://www.repeatmasker.org]; PlantCARE95; TF binding sites unscovered by PlantRegMap were illustrated with R package drawProteins8 (Figs. 5d and 6b). Additional details regarding TFs are provided in the Supplementary Note 8.

Genome duplication and intergenomic comparison for asterids. To investigate genome evolution across asterids, we searched for genome-wide duplications in the order Cornales and Ericales. First, we identified different modes of gene duplication as whole-genome duplicates (WGD), tandem duplicates (TD), proximal duplicates (less than 10 gene distance on the same chromosome: PD), transposed duplicates (transposon-mediated: TRD), or dispersed duplicates (other duplicates than WGD, TD, PD and TRD: DSD) using DupGen_finder49 with default parameters. Second, the Ka (number of substitutions in nonsynonymous site), Ks (number of substitutions in synonymous site), and Ka/Ks values were estimated for gene pairs generated by different modes of duplication based on the TN model in KaKs_calculator (version 2.0)47, after conversion of amino acid alignments into the corresponding codon alignments with PAL2NAL (version 14)47.

For intergenomic comparison, we compared Vitis vinifera genome with five asterids species (Camellia sinensis, Camptotheca acuminata, Rhododendron delavayi, R. williamsianum, R. simsii), and also compared R. simsii genome with that of Actinidia chinensis and a haploptype of Vaccinium corymbosum. The orthologs between these species were identified using MScScanX_jh49. Subsequently, Ks substitution rates of gene pairs across syntenic blocks were calculated. Finally, we illustrated Ks distribution and the dotplots of orthologous blocks using MScScanX_jh49.

WGD time was estimated with asterids/rosids divergence time means (mean: 121 Ma)54 as an age constraint. Ks peaks of V. vinifera vs. five species (Camellia sinensis, Camptotheca acuminata, R. delavayi, R. williamsianum, and R. simsii) syntenic orthologs allows calculating Ks per year (r) following r = Ks/(2 × (divergence time))16. We applied the same r value to calculate the time of WGD events for each species.

Flowering time and flower color genes. We searched for homology of flowering time control genes in A. thaliana by querying FLOR-DSc4, a recently developed database which categorizes 295 protein-coding and 11 miRNA genes and describes their interactions within Arabidopsis’ genome. Flower color is mainly determined by anthocyanins and carotenoids5,33. Given the importance of anthocyanins/carotenoids accumulation in R. simsii, we annotated genes associated with the anthocyanins/carotenoids biosynthesis pathways by querying the Plant Metabolic Network59 and validated by Semi-Automated Validation Infrastructure (version 3.02)59, after enzymatic protein annotations for coding genes through the Ensemble Expression Prediction Pipeline (E2P2) package (version 3.1) (see “URLs” section).

R. simsii transcription factors. We used PlantRegMap60 to identify TFs in the R. simsii genome. Among all identified transcription factors, we further analyzed the phylogenies, protein-covared motifs, and protein structures of MYB, bHLH, WD40, WRKY and ERF. Additional details are available in the Supplementary Note 6 and 7.

Metabolic pathways construction based on TO-GCNs. In order to identify core genes of flower timing and color development, we performed a gene expression study using full transcriptome data. For the gene expression study, 25 frozen corolla tissues from five developmental stages and five times replicates were collected with a mortar and pestle and RNA was isolated using the NEBNext Poly (A) mRNA Magnetic Isolation Module. RNA quality was determined on an Agilent 2100 BioAnalyzer. All of the 25 sequencing libraries were prepared using the NEBNext Ultra RNA Library Prep Kit for Illumina. Next, 150 bp PE mRNA sequencing was performed on an Illumina HiSeq X Ten. These Illumina reads were processed using Trimmomatic (version 0.36)51 and Cutadapt (version 1.31)52, and then mapped to the final assembly using HiSat2 (version 2.1.0). Only uniquely mapped paired-end reads were retained for read counting of the annotated genes by featureCounts42 to generate the count and Transcripts per Kilobase Million (TPM) expression. Differential gene expressions (DGE) analyses between the five stages were performed with DESeq24, with 0.05 as the FDR cut-off and a log2 fold change (FC) cut-off of 1 (Supplementary Fig. 2). We used a recently developed method for reconstructing time-ordered gene co-expression networks (TO-GCNs)59. Significantly differentially expressed genes with average TPM less than 0.5 were excluded. The Pearson correlations between 6,915 azalea and 7,449 genes less than 0.5 or above 0.81 in C1 + C2 + GCN were visualized in graphs using Cytoscape65.

Finally, 2K sequences in the upstream of six core genes (Rhizosmis03G011400, Rhizosmis08G0132500, Rhizosmis10G0146300, Rhizosmis11G0126300, Rhizosmis13G0068400, and Rhizosmis13G0080100) were extracted and queried against PlantRegMap and PlantCARE95. TF binding sites unscovered by PlantRegMap were illustrated with R package drawProteins8 (Figs. 5d and 6b). Additional details regarding TFs are provided in the Supplementary Note 8.

Gene functional enrichment. Hypergeometric tests were performed to determine whether specific functional categories from GO were significantly over-represented in R. simsii gene sets within the genome. Functional enrichment was tested using the R package clusterProfiler (version 3.6.0)48 with the background set of all annotated R. simsii genes in this study.
Received: 11 January 2020; Accepted: 11 September 2020; Published online: 19 October 2020

References

1. Geng, Y. Y. *The genus Rhododendron of China* (Shanghai Scientific and Technical Publications, Shanghai, 2014).

2. Splemmer, H. Ein system der gattung Rhododendron L. Botanische Jahrb. Syst. 74, 511–533 (1949).

3. De Riek, J. et al. Ornamental Crops. 237–271 (Springer, Cham, 2018).

4. Tanaka, Y., Sasaki, N. & Ohiyama, A. Biosynthesis of plant pigments: anthocyanins, betalains and carotenoids. Plant J. 54, 733–749 (2008).

5. Wensing, C. A. & Rausher, M. D. Lessons from the evolution of flower color on targets of selection. J. Exp. Bot. 63, 5741–5749 (2012).

6. Mizuta, D., Ban, T., Miyajima, I., Nakatsuka, A. & Kobayashi, N. Comparison of flower color with anthocyanin composition patterns in evergreen azalea. Sci. Horticult. 122, 594–602 (2009).

7. Soza, V. L. et al. The Rhododendron genome and chromosomal organization provide insights into shared whole genome duplications across the heath family (Ericaceae). Genome Biol. Evol. 11, 3357–3371 (2019).

8. Zhang, L. et al. The draft genome assembly of Rhododendron delavayi Franch. var. delavayi. Gigasscience 6, 1–11 (2017).

9. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).

10. Ross, M. G. et al. Characterizing and measuring bias in sequence data. Genome Biol. 14, R51 (2013).

11. Chen, R. Chromosome atlas of major economic plants genome in China. (Science Press, Beijing, 2003).

12. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3215 (2015).

13. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Res. 46, e126 (2018).

14. Jaillon, O. et al. The grapevine sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449, 463–467 (2007).

15. Ivezic, M. et al. A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. Nat. Genet. 48, 657–666 (2016).

16. Badouin, H. et al. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature 546, 148–152 (2017).

17. Chase, M. W. et al. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot. J. Linn. Soc. 181, 1–20 (2016).

18. Colla, M. et al. Haplotype-phase genomics and evolution of pythionutrient pathways of tetraploid blueberry. Gigasience 8, gz012 (2019).

19. Pilkington, S. M. et al. A manually annotated Actinidia chinensis var. chinensis (kiwifruit) genome highlights the challenges associated with draft genomes and gene prediction in plants. BMC Genomics 19, 257 (2018).

20. Wei, C. et al. Draft genome sequence of Camellia sinensis var. sinensis provides insights into the evolution of the tea genome and tea quality. Proc. Natt Acad. Sci. USA 115, E4151–E4158 (2018).

21. Zhao, D. et al. De novo genome assembly of Camptotheca acuminata, a natural source of the anti-cancer compound camptothecin. Gigasience 6, 1–7 (2017).

22. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomes. Genome Biol. 20, 238 (2019).

23. De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. Bioinformatics 22, 1269–1271 (2006).
Acknowledgements

This work was supported by grants from the Second Tibetan Plateau Scientific Expedition and Research (STEP) program (2019QZKK0502), Project of Construction of World Class Universities in Beijing Forestry University (2019XKJ0030), the Fundamental Research Funds for the Central Universities in Beijing Forestry University (2018LCLB08), National Key R&D Program of China (2017YFA0605100), and Strategic Priority Research Program of the Chinese Academy of Sciences (XDA23080000).

Author contributions

J.F.M., F.S.Y., and Y.V.P. conceived and designed the study; F.S.Y., S.N., T.L.S., X.C.T., H.L., S.S.Z., Y.T.B., K.H.J., J.F.G., W.Z., N.A, R.G.Z., Q.Z.Y., X.Z.W., C.M., and N.R.S. prepared the materials, conducted the experiments, analyzed data and prepared results; J.F.M., F.S.Y., and S.N. wrote the manuscript; X.R.W., I.P., Y.A.E., and Y.V.P. were involved in data interpretation and finalizing the manuscript draft. All authors read and approved the final draft.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-18771-4.

Correspondence

and requests for materials should be addressed to Y.V.P. or J.F.M.

Peer review information

Nature Communications thanks Wen-Hsiung Li, Marek Mutwil, Robert VanBuren and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information

is available at http://www.nature.com/reprints

Publisher’s note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020