

Phenotypic and genomic signatures across wild *Rosa* species open new horizons for modern rose breeding

In the format provided by the
authors and unedited

Supplementary Notes: Revisions of *Rosa* species in *Flora of China* for this study

1. Newly added: Newly published *Rosa* species and varieties, that were not recorded in the *Flora of China*, but are included in this study.

1.1 *R. tomurensis* (sect. *Rosa*)

This species was first considered as a variety of *R. laxa* and was recorded in *Rosa in Xinjiang* (Liu and Cong, 2020). After years of investigation, it was confirmed and published as an independent species in *Rosa*, and it was officially named *R. tomurensis* L. Luo, C. Yu & Q. X. Zhang (Deng et al., 2022).

1.2 *R. cinnamomea* (sect. *Rosa*)

This species is the exemplary type species of *Rosa*, but was not recorded in the *Flora of China*. Field investigations have revealed a wild distribution of *R. cinnamomea* in the north of Xinjiang, China, which is why we included this species in our study.

1.3 *R. yangii* (sect. *Chinenses*)

This species was officially published as a new species from Yunnan, China (Luo et al., 2024; Lyu et al., 2023), and is considered to be one of the important wild germplasms for studying the evolution and traits of sect. *Chinenses*.

1.4 *R. funingensis* (sect. *Synstylae*)

R. funingensis and its variant *R. funingensis* f. *rosea* were both collected from Yunnan Province, China. They were officially documented and published as new species of *Rosa* (Zheng et al., 2023).

1.5 *R. forrestiana* var. *maculata* (sect. *Rosa*)

This new variety of *Rosa* was collected from northwest Yunnan, China (Tang et al., 2024). It was distinguished by its purple spots at the petal base, a rare character in wild *Rosa* resources.

2. Reclassification: Based on phenotype examination and specimen comparison, combined with previous research and phylogenetic results of this study, it is suggested to revise the botanical sections of certain materials.

2.1 *R. tsinglingensis* and *R. koreana* (from sect. *Pimpinellifoliae* to sect. *Rosa*)

The main characteristics defined by sect. *Pimpinellifoliae* are flower solitary, without bracts. In specimen comparison and phenotypic identification, both *R. tsinglingensis* and *R. koreana* have 1-3 flowers with bracts.

Molecular evidence obtained in this study showed that both *R. tsinglingensis* and *R. koreana* were clustered into sect. *Rosa*, based on phylogenetic trees. This result is consistent with previous phylogenetic studies of *Rosa*, where Debray et al. (2022) showed that *R. tsinglingensis* was clustered with other species in sect. *Rosa*. The results of a study by Fougère-Danezan et al. (2015) showed that both species clustered in sect. *Rosa*.

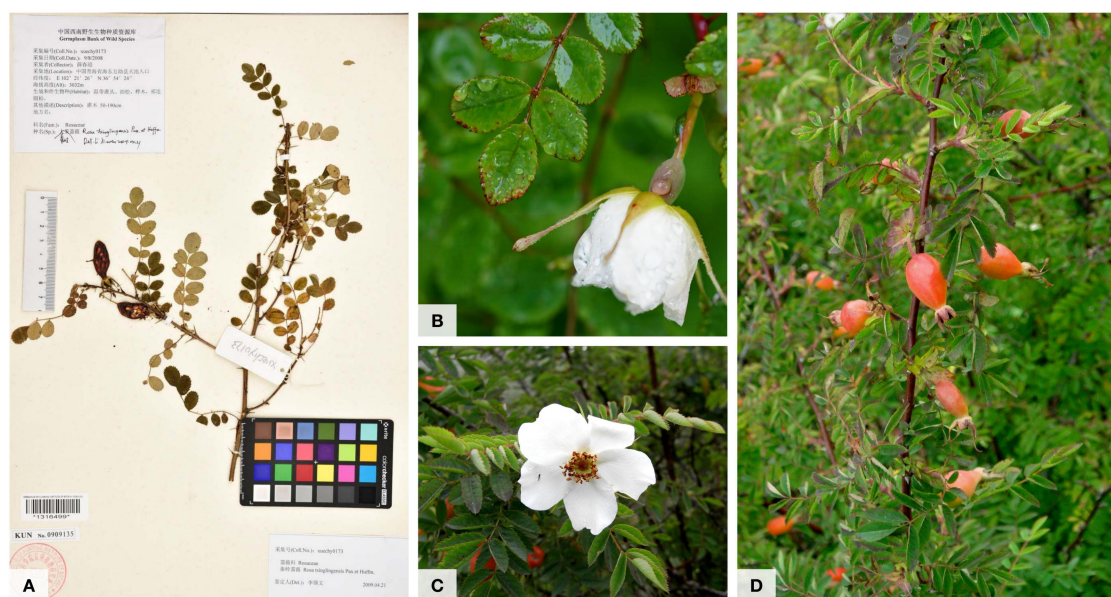


Fig. 2.1a. *R. tsinglingensis*. **A.** Specimen of *R. tsinglingensis**; **B-C.** Flower; **D.** Hips.

*Specimen downloaded from Chinese Virtual Herbarium (<https://www.cvh.ac.cn/index.php>), Barcode: 1316499.

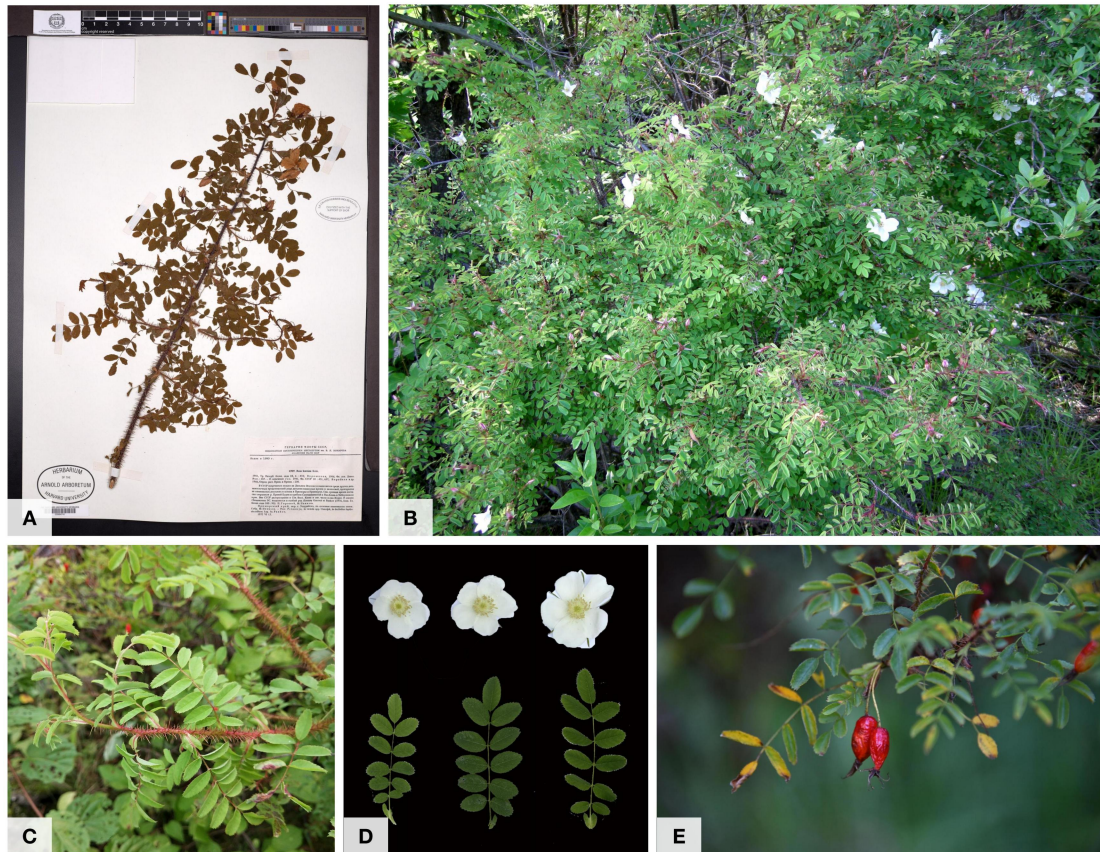


Fig. 2.1b. *R. koreana*. **A.** Specimen of *R. koreana** ; **B.** Whole plant; **C.** Branches and prickles; **D.** Flowers and leaves; **E.** Hips.

*Specimen downloaded from Harvard University Herbaria & Libraries (<https://kiki.huh.harvard.edu>), Barcode: 00984741.

2.2 *R. pseudobanksiae* (from sect. *Rosa* to sect. *Banksianae*)

In terms of morphology, this species differs substantially from sect. *Rosa*, especially in terms of fruit morphology, which is more similar to sect. *Banksianae* and sect. *Synstylae*.

Molecular evidence suggests that this species has a hybrid origin between *R. banksiae* var. *normalis* and *R. multiflora* var. *cathayensis*. The phylogenetic analysis in this study categorized this species into sect. *Banksianae*, and studies by Zhang et al. (2020) and Guan (2023) also support this revision.



Fig. 2.2. *R. pseudobanksiae*. **A.** Specimen of *R. pseudobanksiae** ; **B.** Flowers; **C.** Hips; **D.** Leaves; **E.** Branch and prickles.

*Specimen downloaded from Chinese Virtual Herbarium (<https://www.cvh.ac.cn/index.php>), Barcode: 00020710.

2.3 *R. praelucens* (from sect. *Microphyllae* to sect. *Rosa*)

The geographical distribution and phylogenetic relationship of *R. praelucens* are distant from *R. roxburghii* and *R. kweichowensis*, which belong to sect. *Microphyllae*.

Qiu et al. (2012), Fougère-Danezan et al. (2015), Deng et al. (2016), and Debray et al. (2022) have conducted research based on different sampling systems. They all came to same conclusion, namely that *R. praelucens* should be included in sect. *Rosa*. The phylogenetic trees generated in this study also support this classification.



Fig. 2.3. *R. praelucens*. **A.** Flowers; **B.** Leaf; **C-D.** Hips.

3. Re-identification: Based on phenotype examination and specimen comparison, combined with previous research and the phylogenetic results of this study, several species are suggested to be combined and discussed together.

3.1 *R. lichiangensis* (variety of *R. multiflora*, same species as *R. multiflora* var. *cathayensis*)

The characters of *R. lichiangensis* recorded in the *Flora of China* conform to the pink petal variation of wild roses, which is a characteristic of *R. multiflora* var. *cathayensis*.

A previous study suggested that *R. lichiangensis* had a hybrid origin of *R. soulieana* and *R. multiflora* var. *cathayensis* (Zhu et al., 2015). The phylogenetic analyses conducted in our study show that *R. lichiangensis* is closely related to *R. multiflora* and its cultivated samples, indicating that it should be considered as a variety rather than an independent species. Therefore, we suggest to treat *R. lichiangensis* as a variant of *R. multiflora*, the same species as *R. multiflora* var. *cathayensis*.

3.2 *R. henryi* (same species as *R. sambucina*)

The characteristics of *R. henryi* and *R. sambucina* recorded in the *Flora of China* are identical, and the two species may be synonymous. As the name *R. sambucina* was published before *R. henryi*, it is suggested to combine both species and use *R. sambucina* as the official name. Such a revision is also supported by the phylogenetic results obtained in this study.



Fig. 3.2. *R. sambucina* and *R. henryi*. **A-B.** Flowers and hips of *R. sambucina*; **C-D.** Flowers and hips of *R. henryi*.

3.3 *R. weisiensis* and *R. degenensis* (same species as *R. duplicata*)

The leaflets of these three species are all double serrated, showing only slight differences in the size of leaflets. The flowers and fruits characteristics are also the same. Based on field examinations, the geographical distribution of these three species overlaps.

Three samples were clustered into the same group both in this study and in a previous phylogenetic study (Zhu et al., 2015). All this evidence suggests that the three species are actually the same species with different names, and should be studied as one species, or as a composite group.



Fig. 3.3a. *R. weisiensis*. **A.** Specimen of *R. weisiensis**; **B.** Flowers; **C.** Hips.

*Specimen downloaded from Chinese Virtual Herbarium (<https://www.cvh.ac.cn/index.php>), Barcode: 00020713.



Fig. 3.3b. *R. degenensis*. **A.** Specimen of *R. degenensis**; **B.** Flowers; **C.** Hips.

*Specimen downloaded from Chinese Virtual Herbarium (<https://www.cvh.ac.cn/index.php>), Barcode: 00061335.



Fig. 3.3c. *R. duplicata*. **A.** Specimen of *R. duplicata**; **B.** Flowers; **C.** Hips.

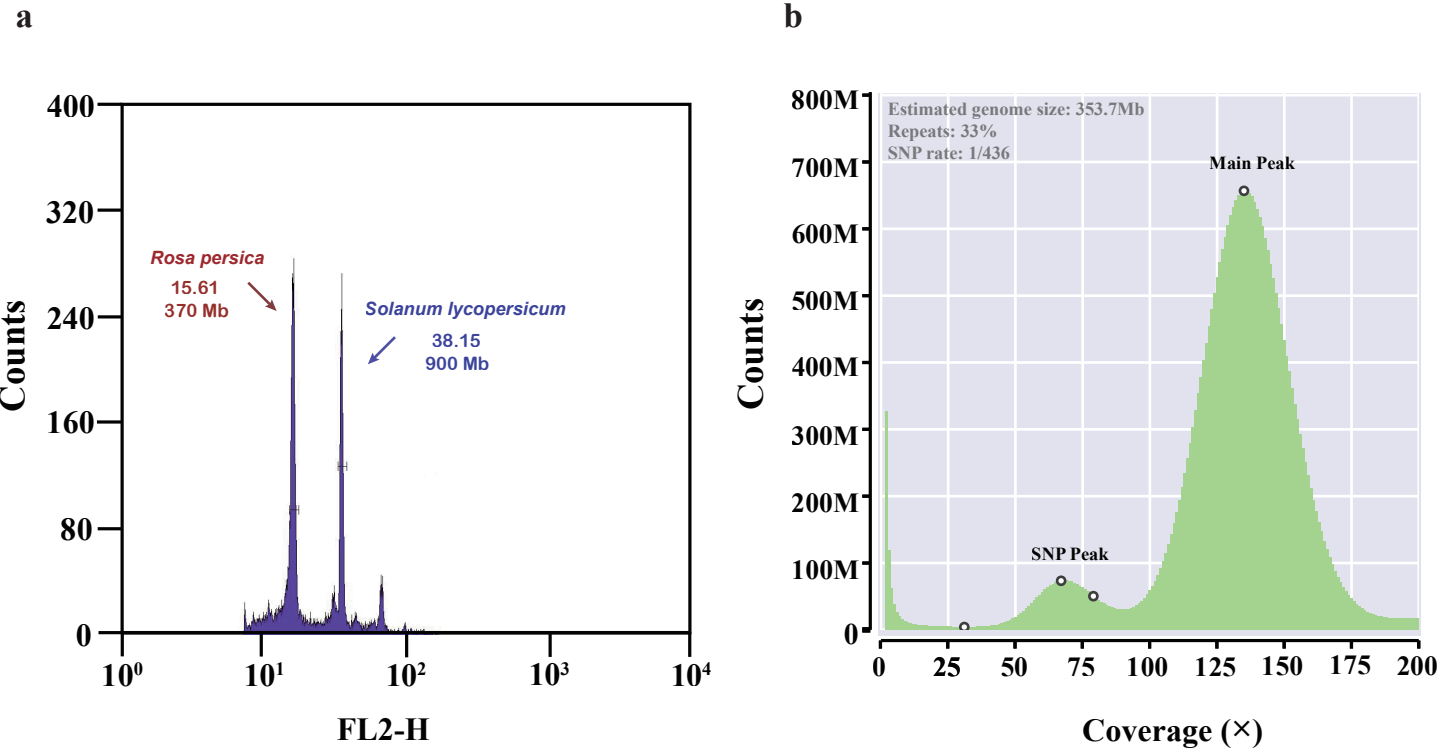
*Specimen downloaded from Chinese Virtual Herbarium (<https://www.cvh.ac.cn/index.php>), Barcode: CDBI0237162.

References:

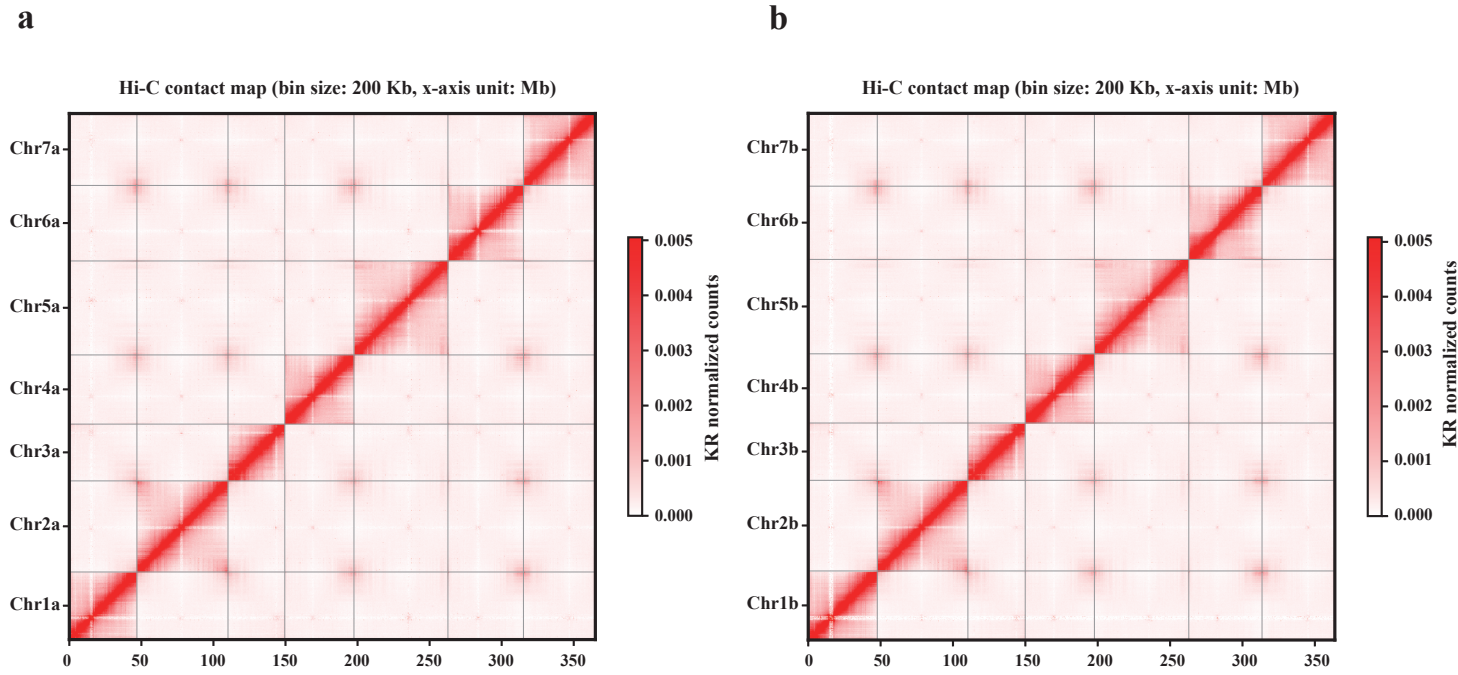
- 1 Liu, S.X. and Cong, Z.F. *Rosa* in Xinjiang. Xinjiang Science and Technology Health Publishing House, Xinjiang, pp. 3–47. (2000) (In Chinese).
- 2 Deng, T. et al. *Rosa tomurensis*, a new species of *Rosa* (Rosaceae) from China. *Phytotaxa* 556, 169-177, doi:10.11646/phytotaxa.556.2.6 (2022).
- 3 Luo, L. Y., Yu-yong. Zhang, Qi-xiang. Genus *Rosa* L. in China., (China Forestry Publishing House, 2024).
- 4 Lyu, P. et al. *Rosa yangii* (Rosaceae), a new species from China. *Kew Bulletin* 78, 663-671, doi:10.1007/s12225-023-10135-7 (2023).
- 5 Zheng, L.-N. et al. *Rosa funingensis* (Rosaceae), a new species from Yunnan, China. *PhytoKeys* 229, doi:10.3897/phytokeys.229.101052 (2023).
- 6 Yu-wei Tang, L. L., Xue-sen Liu, Xiao-ling Tian, Chao Yu, Qi-xiang Zhang & Yu-yong Yang. *Rosa forrestiana* var. *maculata*, a new variety of *Rosa* (Rosaceae) from Yunnan, China. (2024).
- 7 Debray, K. et al. Unveiling the Patterns of Reticulated Evolutionary Processes with Phylogenomics: Hybridization and Polyploidy in the Genus *Rosa*. *Syst Biol* 71, 547-569, doi:10.1093/sysbio/syab064 (2022).

- 8 Fougere-Danezan, M., Joly, S., Bruneau, A., Gao, X. F. & Zhang, L. B. Phylogeny and biogeography of wild roses with specific attention to polyploids. *Annals of Botany* 115, 275-291, doi:10.1093/aob/mcu245 (2015).
- 9 Guan, Y. Study on the chloroplast genome and systematic relationships of the genus *Rosa*. Yunnan Normal University. (2023) (in Chinese).
- 10 Qiu, X.-Q, Zhang, H., Wang, Q. -G, et al. Phylogenetic relationships of wild roses in China based on nrDNA and matK data[J]. *Scientia Horticulturae*, 140: 45-51. (2012).

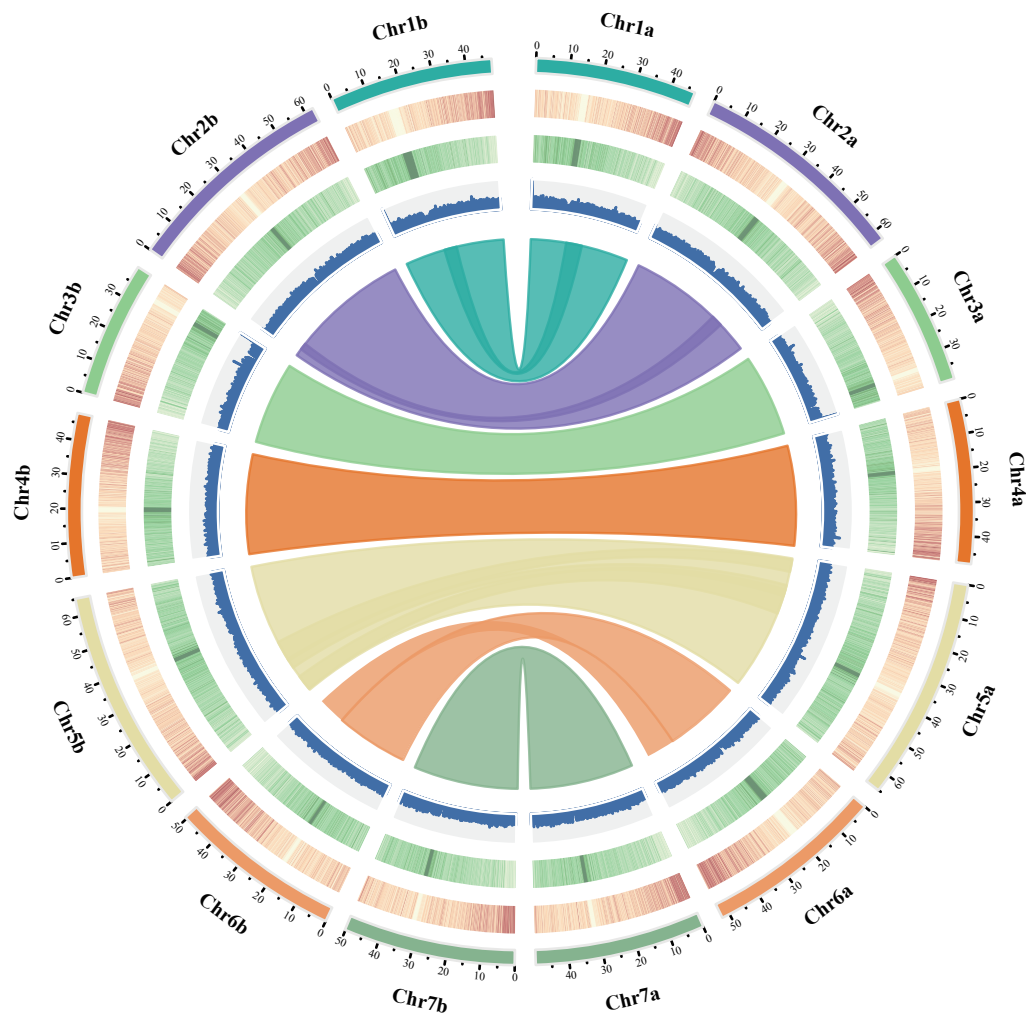
Supplementary Figures



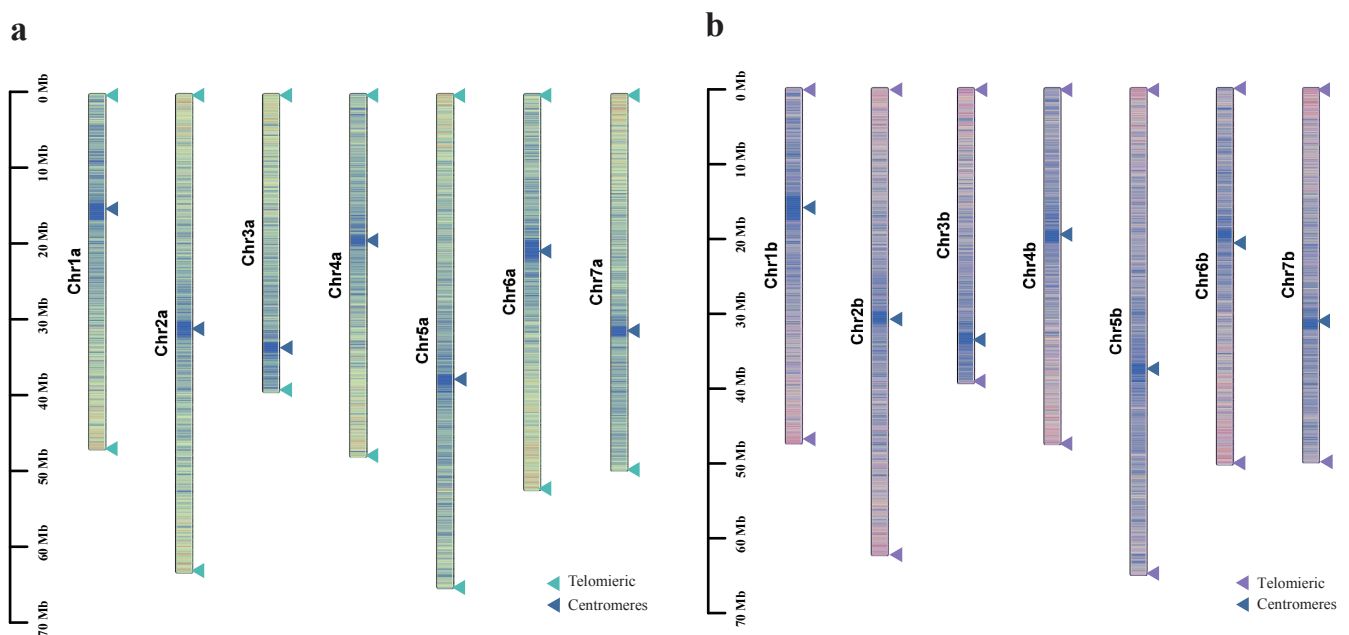
Supplementary Fig. 1. Genome size estimation of *Rosa persica*. **a.** C-value estimation of *R. persica* by flow cytometry using *Solanum pimpinellifolium* as the internal reference standards; **b.** K-mer analyses of sequencing reads of *R. persica*.



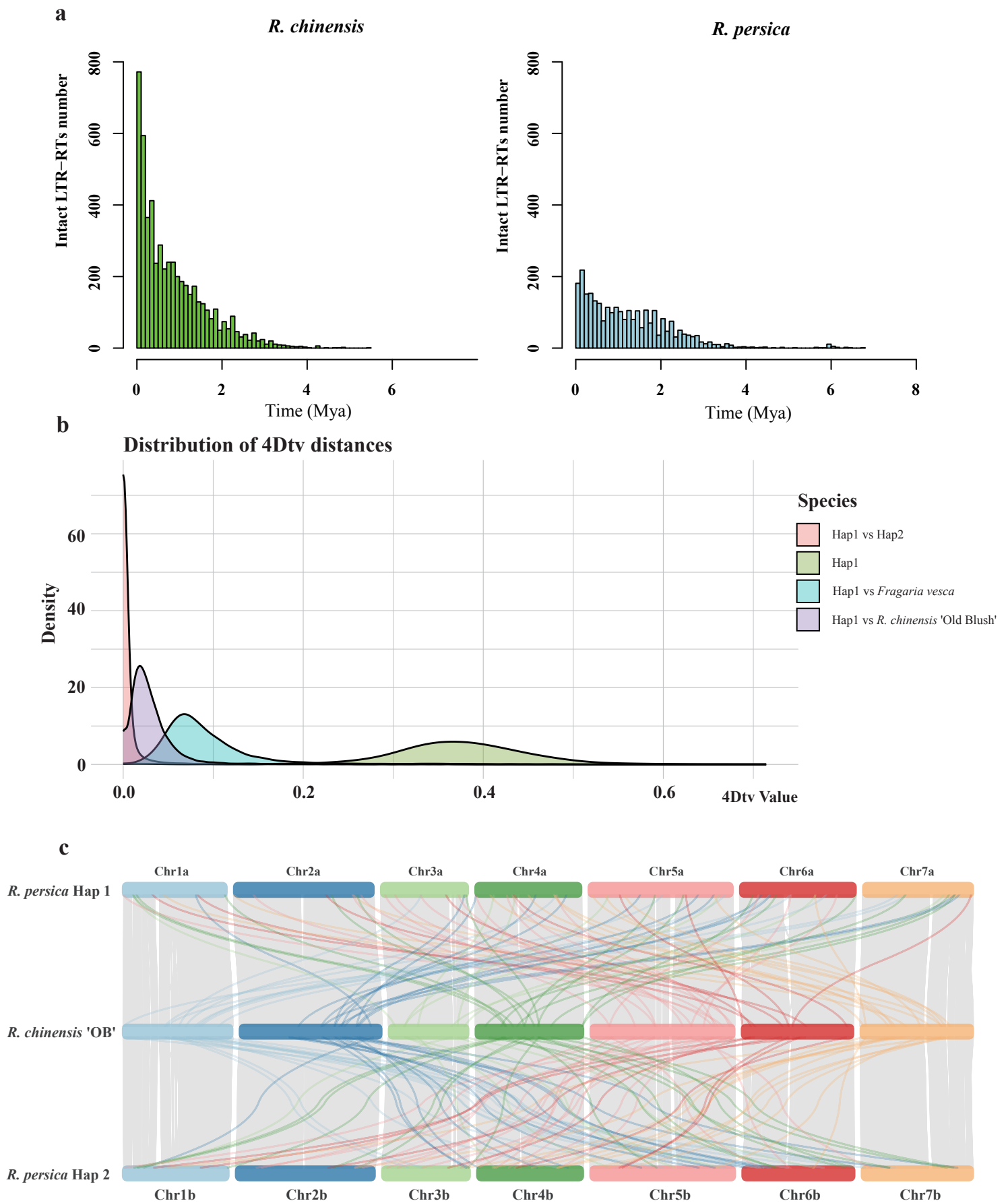
Supplementary Fig. 2. Hi-C interaction heatmap of the *R. persica* genome. **a.** *R. persica* Hap 1; **b.** *R. persica* Hap 2.



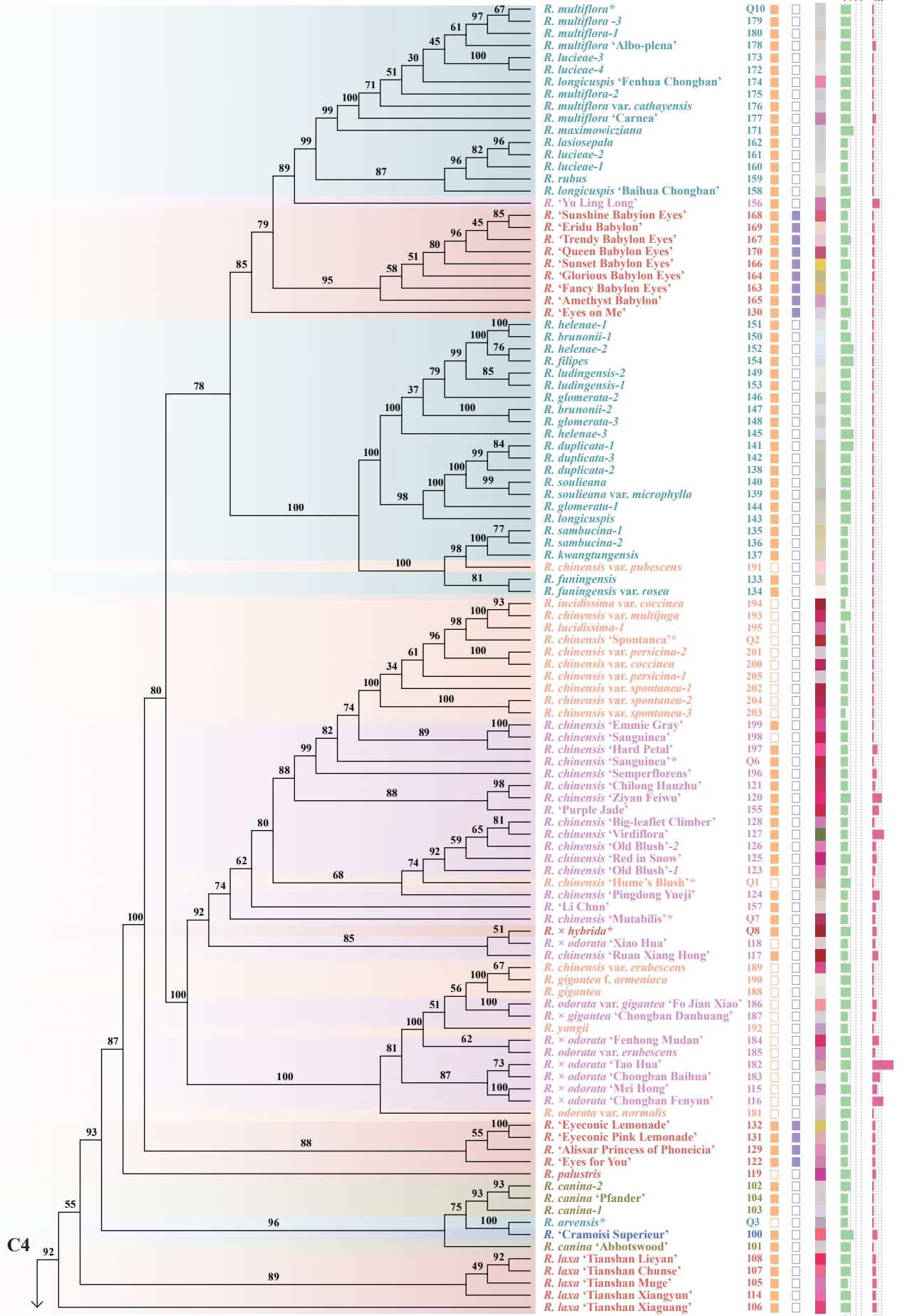
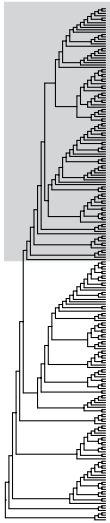
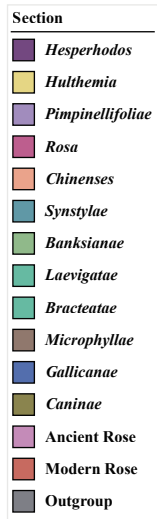
Supplementary Fig. 3. Circos plot of the *Rosa persica* genome. The circles from the inner circle to the outer circle represent: Collinearity; InDel density, SNP density, GC content, Repeat content, Gene density, Chromosome. Window size is 100 kb.



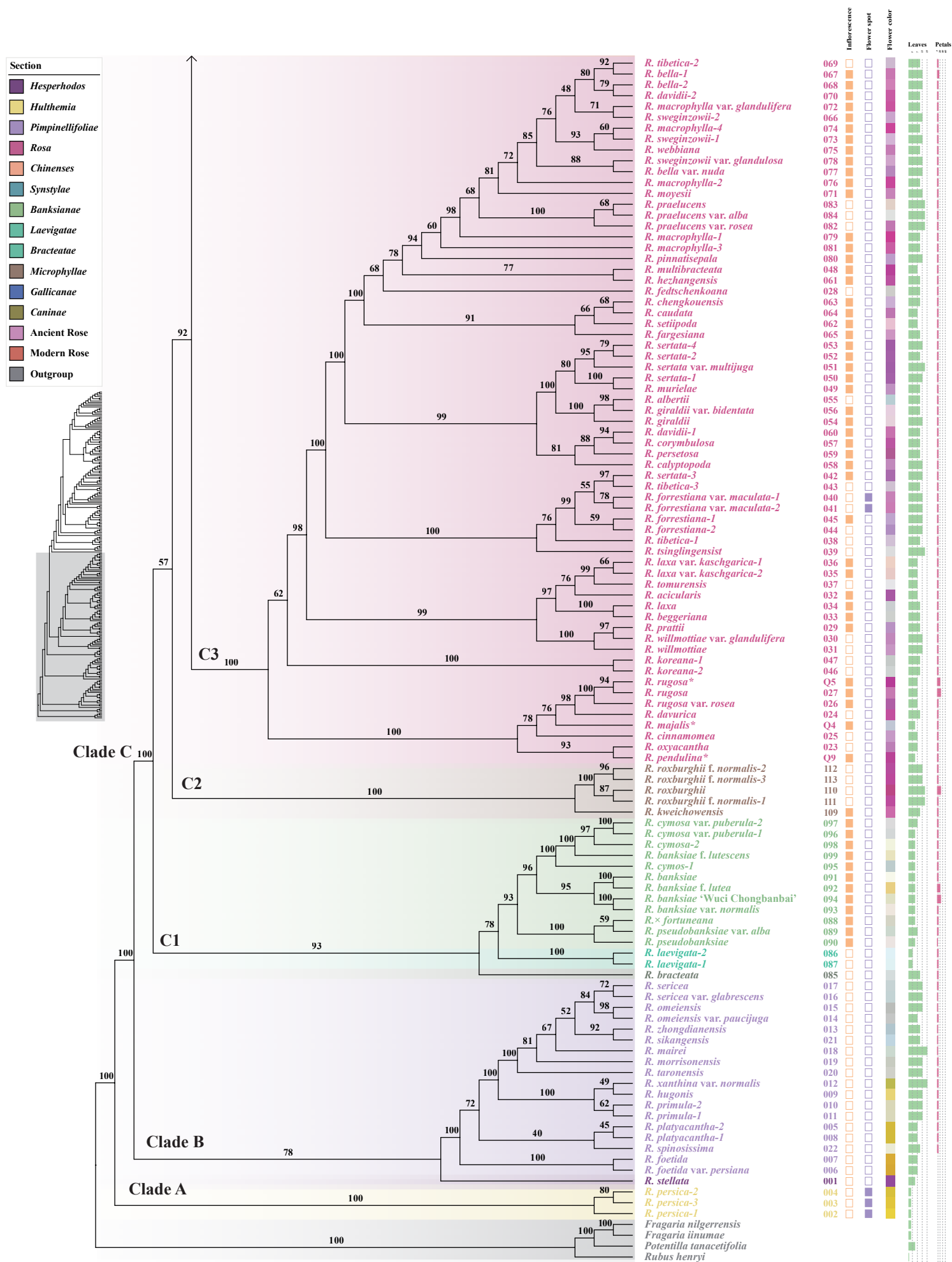
Supplementary Fig. 4. Positions of centromeres and telomeres of two haplotypes. a. Seven chromosomes of *R. persica* Hap 1; **b.** Seven chromosomes *R. persica* Hap 2. The heat map represents gene density.



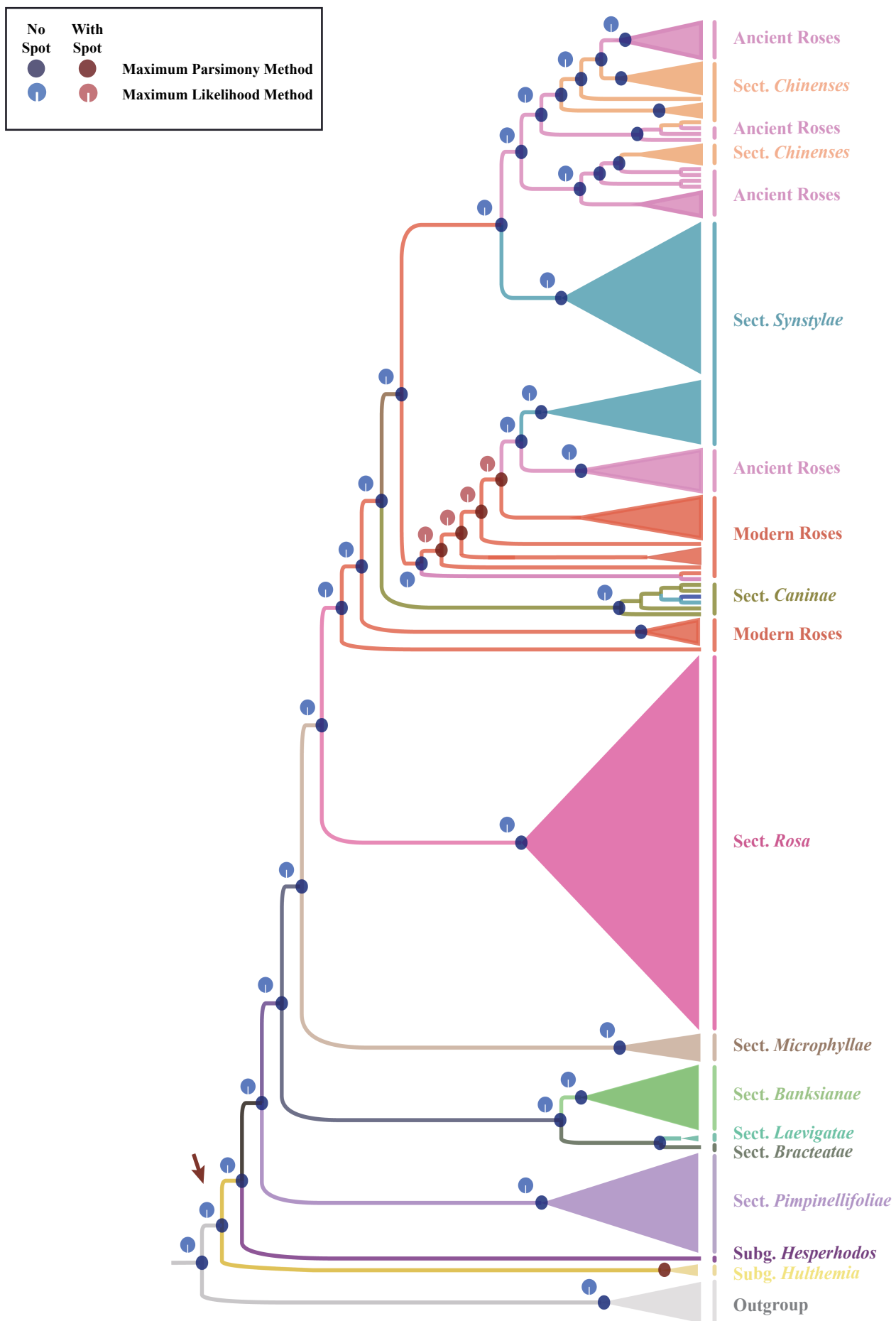
Supplementary Fig. 5. Comparative genome analysis of *R. persica*. **a.** LTR insertion time of *R. chinensis* 'Old Blush' and *R. persica*. **b.** Whole-genome duplication analysis of *R. persica*. **c.** Syntenic relationship between the *R. persica* genome and the *R. chinensis* 'Old Blush' genome.



Supplementary Fig. 6. Phylogenetic tree of 215 *Rosa* accessions estimated by Maximum Likelihood algorithm using IQ-TREE, based on 6,048 single-copy SNPs (continued). The bootstraps are indicated above the branch. The colors represent accessions from different botanical groups. *Fragaria nilgerrensis*, *Fragaria iinumae*, *Potentilla tanacetifolia*, and *Rubus henryi* were used as outgroup. Trait information for each accession is listed on the right. Sample IDs correspond to Supplementary Table 13.

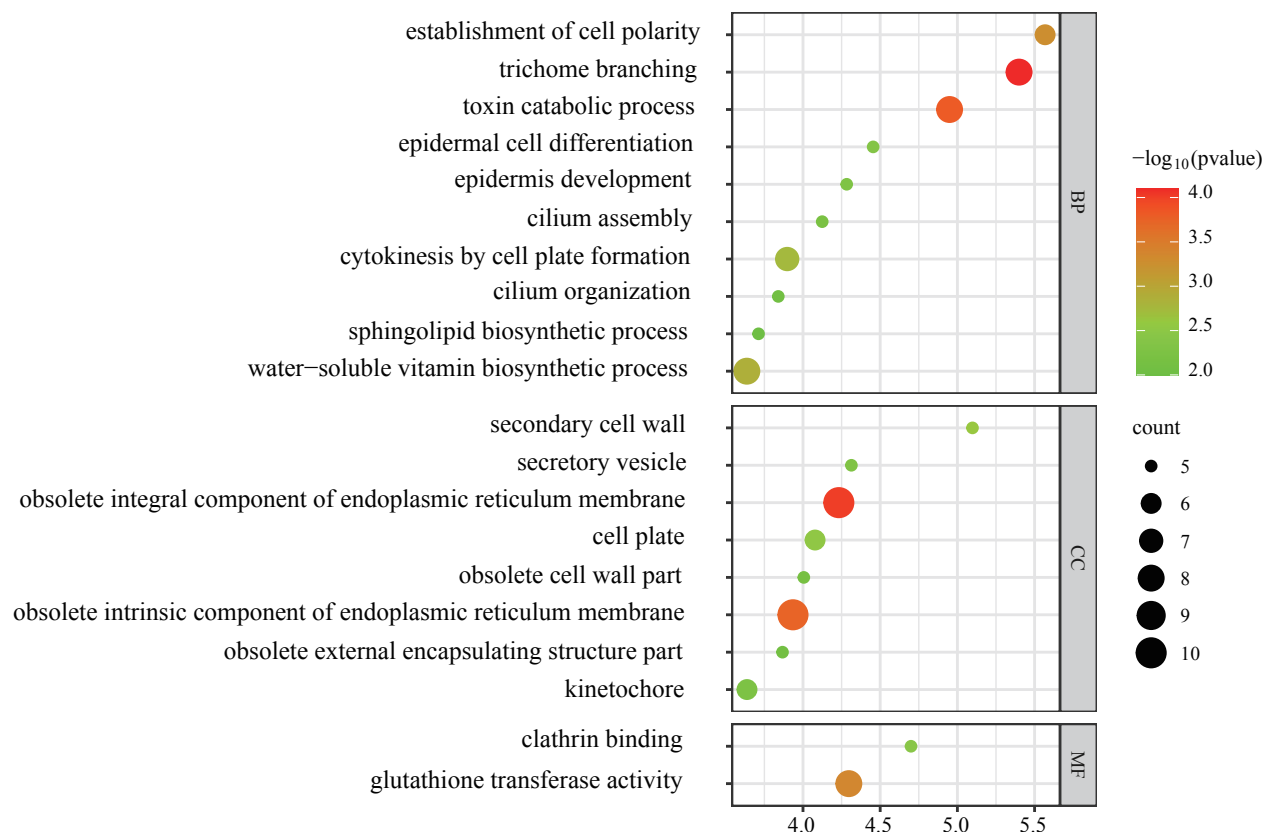


Supplementary Fig. 6. Phylogenetic tree of 215 *Rosa* accessions estimated by Maximum Likelihood algorithm using IQ-TREE, based on 6,048 single-copy SNPs. The bootstraps are indicated above the branch. The colors represent accessions from different botanical groups. *Fragaria nilgerrensis*, *Fragaria iinumae*, *Potentilla tanacetifolia*, and *Rubus henryi* were used as outgroup. Trait information for each accession is listed on the right. Sample IDs correspond to Supplementary Table 13.

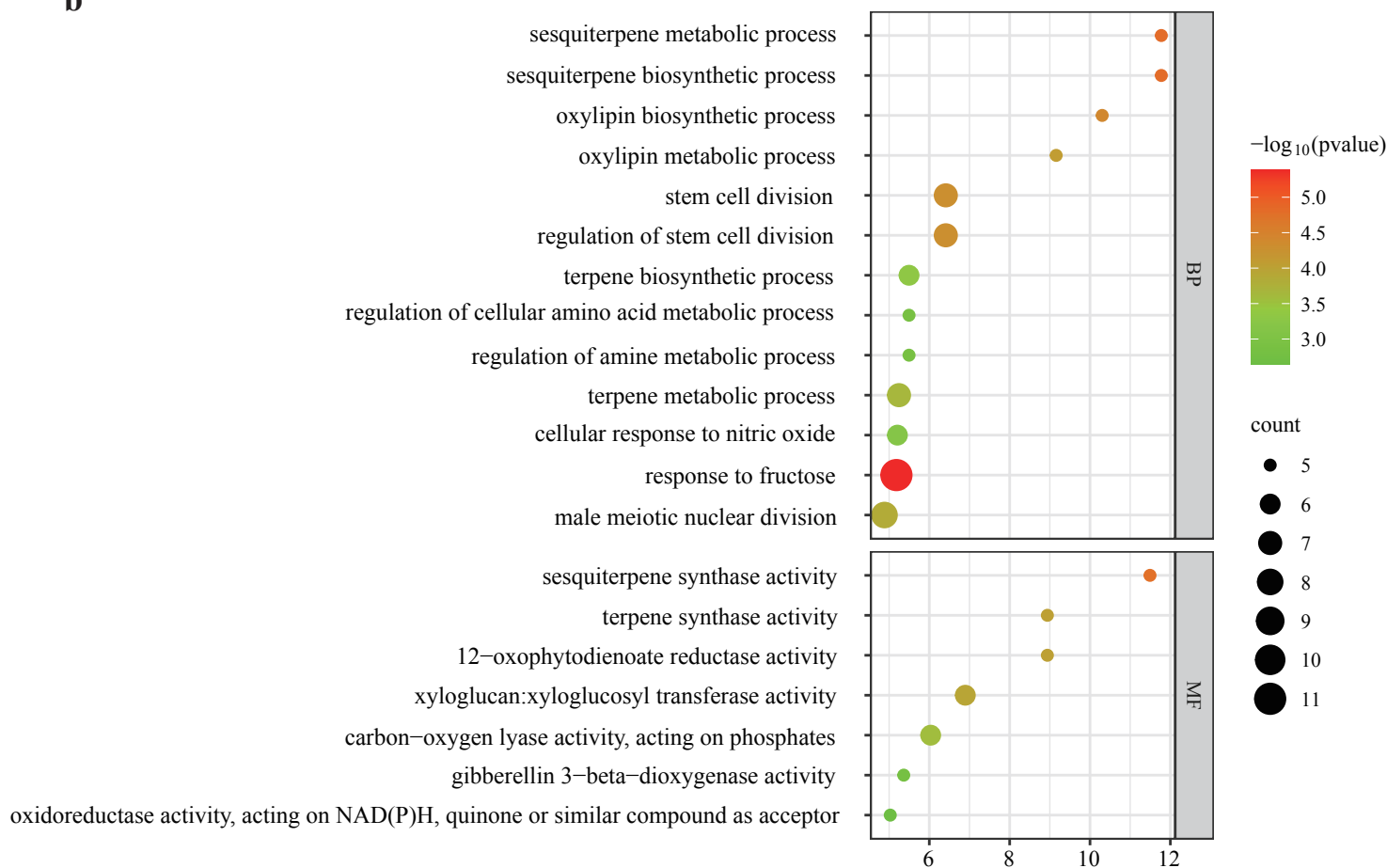


Supplementary Fig. 7. Ancestral state reconstruction for petal spot of *Rosa*, performed using both the maximum parsimony method (at the node of the tree) and the maximum likelihood method (above the node of the tree) for the backbone of nuclear phylogeny. The color keys and pie diagrams in the internal nodes represent the most likely ancestral character states and the relative probabilities of each alternative state. The red arrows point to the common ancestor nodes of *Rosa*, showing the inferred ancestral character states of petal spot.

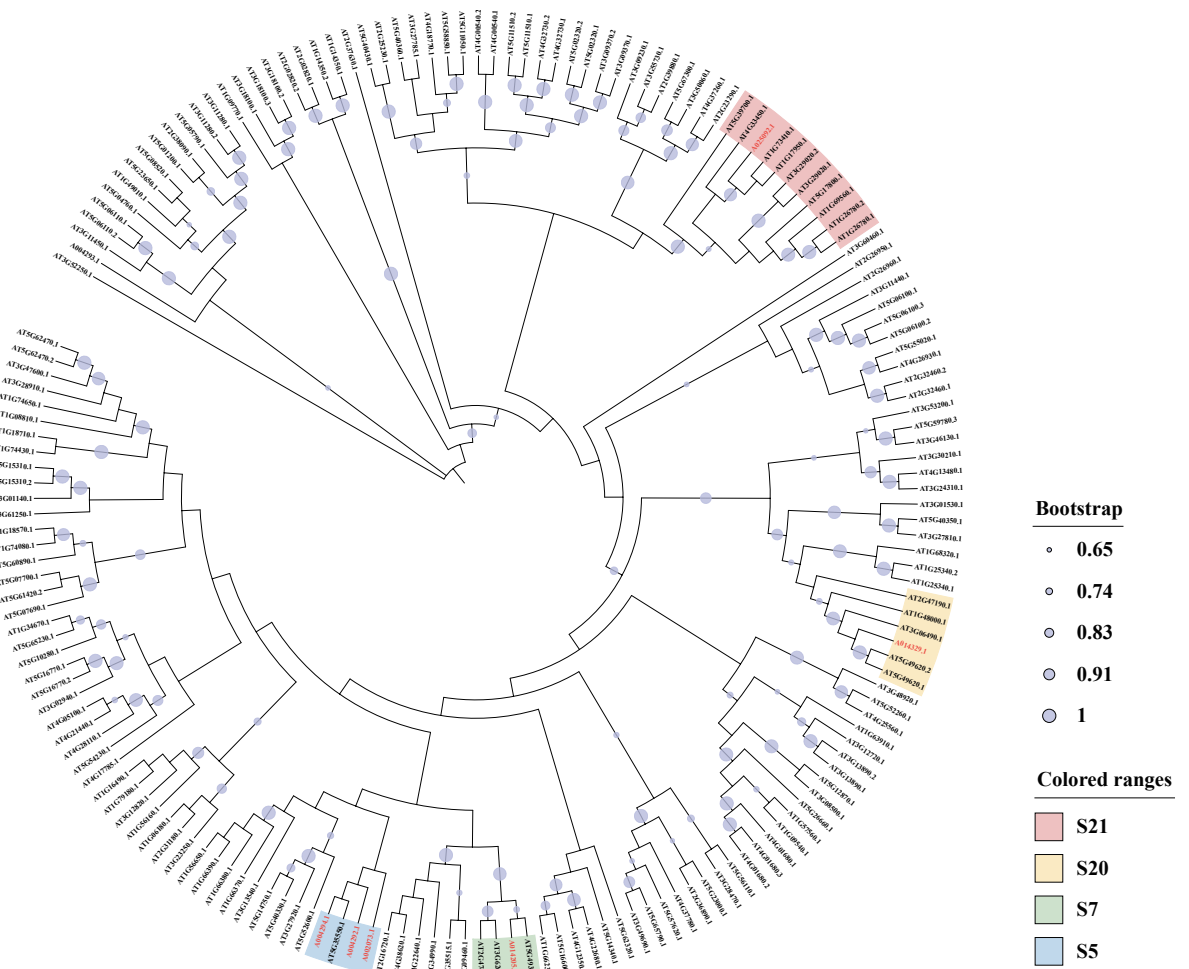
a



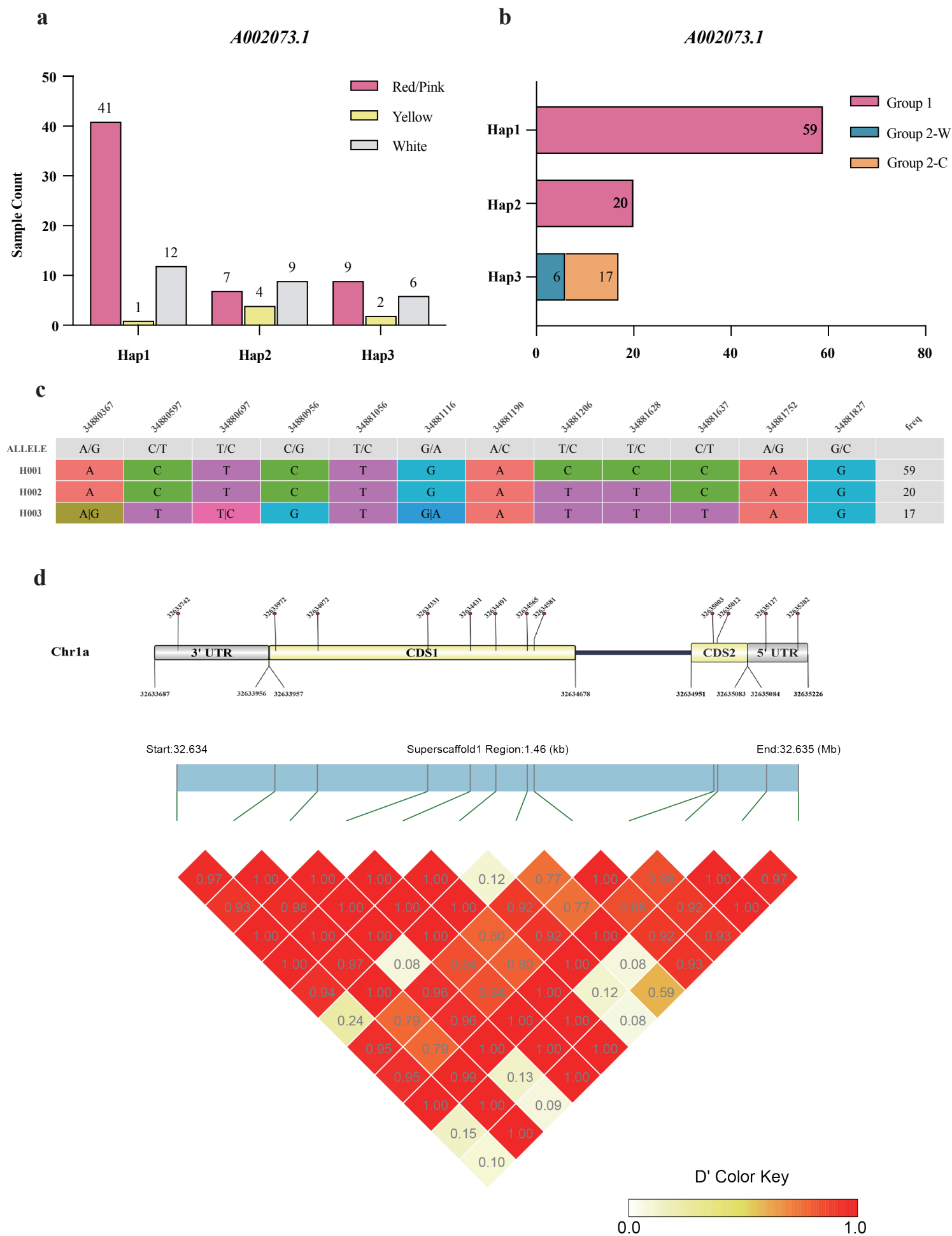
b



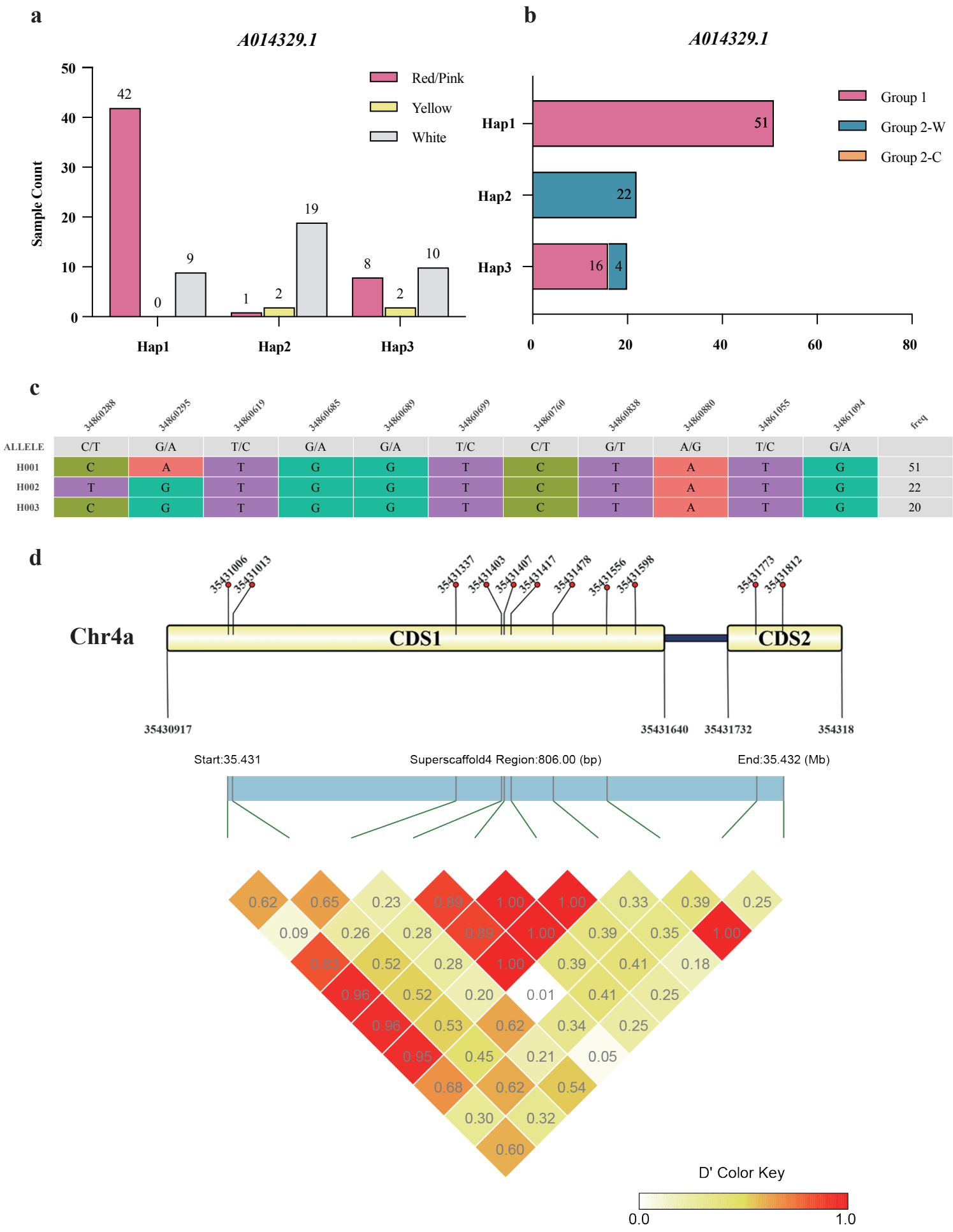
Supplementary Fig. 8. Gene Ontology enrichment of a. candidate genes in selective sweeps contrast Group 1 and Group 2 populations; **b.** candidate genes in selective sweeps contrast Group 2-W and Group 2-C populations.



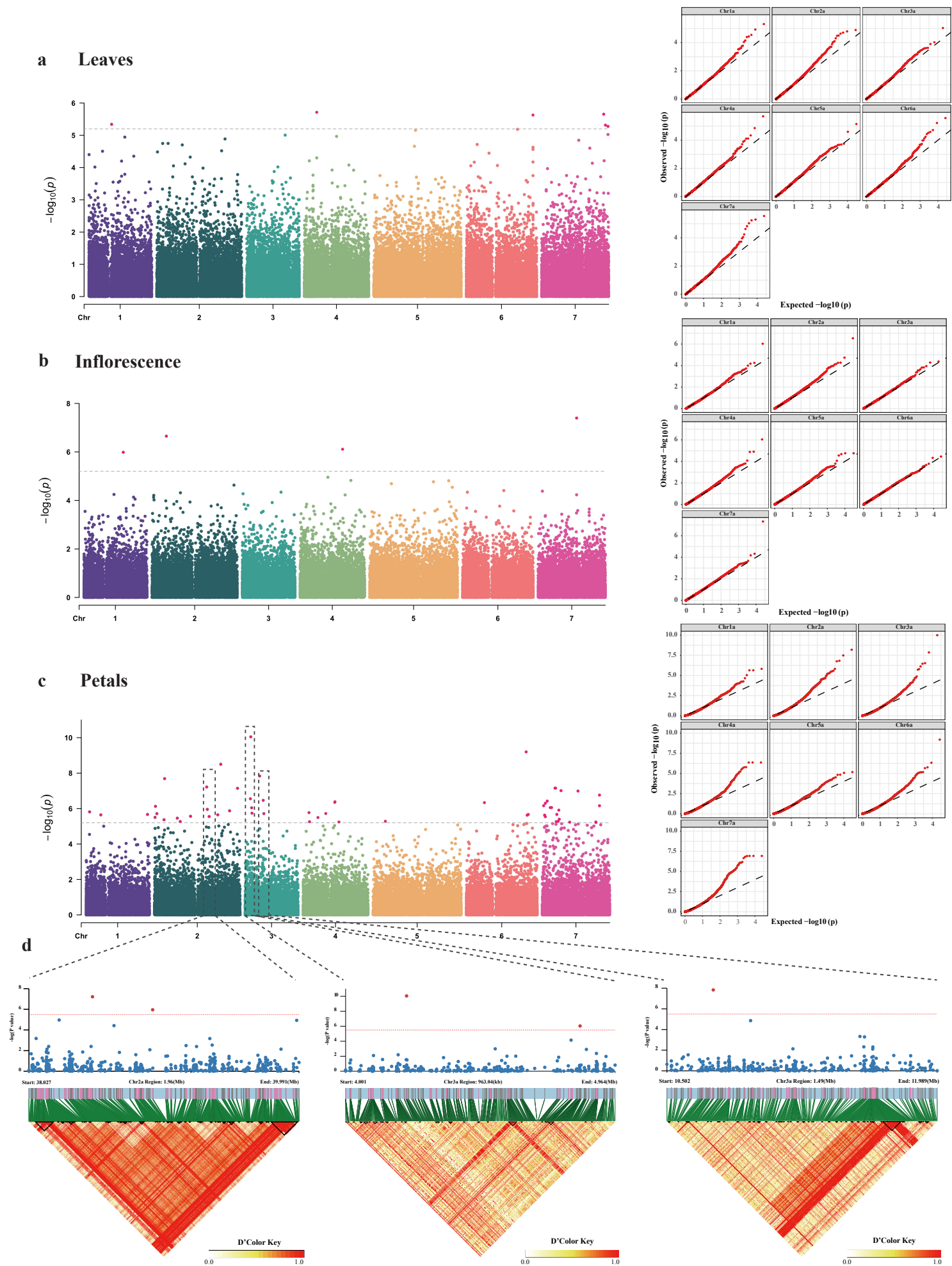
Supplementary Fig. 9. Neighbour-joining phylogenetic tree constructed using sequences of the MYB transcription factor family in *Arabidopsis*, and genes identified in selective sweeps in this study (marked red).



Supplementary Fig. 10. Haplotype analyses of *A002073.1*. **a.** Distribution of the flower color phenotypes (Red/pink, Yellow, and White) corresponding to the different haplotypes (Hap1-3); **b.** Haplotypes of *A002073.1* corresponding to accessions from different groups; **c.** Haplotype classifications of *A002073.1*, each line represent a haplotype and corresponded loci, the last column indicates haplotype frequency; **d.** Gene structure and haplotype heatmap of *A002073.1*.



Supplementary Fig. 11. Haplotype analyses of *A014329.1*. **a.** Distribution of the flower color phenotypes (Red/pink, Yellow, and White) corresponding to the different haplotypes (Hap1-3); **b.** Haplotypes of *A014329.1* corresponding to accessions from different groups; **c.** Haplotype classifications of *A014329.1*, each line represent a haplotype and corresponded loci, the last column indicates haplotype frequency; **d.** Gene structure and haplotype heatmap of *A014329.1*.



Supplementary Fig. 12. Genome-wide association study (GWAS) of flower and leaf traits across *Rosa* populations. The red dotted line denotes the genome-wide significance threshold of 5.2, as determined using a Bonferroni test.

a. Manhattan plot and Q-Q plot of GWAS on the number of leaflets; **b.** Manhattan plot and Q-Q plot of GWAS on inflorescence types; **c.** Manhattan plot and Q-Q plot of GWAS on the number of petals; **d.** Linkage disequilibrium heatmaps for petal number in *Rosa* populations.

Supplementary Methods

Phenotypic trait measurement

Based on comparison with a large number of specimens and consultation with experts and colleagues, a total of 215 *Rosa* accessions were accurately identified and used for this study. The phenotypic traits were investigated in both the field and germplasm gardens. Measurements for all traits were taken from three biological replicates. Normal leaves (compound leaves on vegetative shoots, with at least 3 leaflets) were selected for measurement of the number of leaflets. A total of 3–5 normal leaves per plant were used for the count. The data were also cross-checked with herbarium specimens and information from the *Flora of China*. The inflorescence type was treated it as a binary trait for statistical analysis. Primary branches were used to record the inflorescence types. Flowers were classified as either solitary or in an inflorescence after combining information from photographs, herbarium specimens, and the recorded information in *Flora of China*. The RGB values were extracted from the photographs of each sample in three technical replicates, with the central spots of the petals representing the flower color.

Genome size evaluation and karyotype analysis

The 57.67 Gb whole-genome data from the Illumina HiSeq X platform was used to evaluate the genome size. The potential genome size was estimated by K-mer analysis (K = 19) using JCVI¹ and flow cytometry. For flow cytometry analyses, fresh leaves were collected and chopped, then put into dissociation solution for 10 min, followed by filtering to collect nucleus cells. The collected cells were combined with propidium iodide and RNAase solution, and placed on ice in the dark for 0.5–1 h. The nuclei suspension was analyzed using BD FACS Calibur flow cytometry. The genome size was estimated using *Solanum pimpinellifolium* (~900 Mb) as internal reference standard (Table S1).

Karyotype analysis was carried out on 108 accessions following the method described by Yu *et al.*². Shoot tips were collected from samples and were treated with 0.002 mol/L 8-hydroxyquinoline for 50 min, followed by fixing for 24 h in fresh Farmer's fixative (1:3 glacial acetic acid: absolute alcohol) at 4 °C. After softening with 1 M hydrochloric acid at 55 °C for 10 min, the shoot tips were stained for 15–20 min with carbofuchsin. Chromosome numbers were counted under observation using LEICA ICC50 HD.

Detection of telomeres and centromeres

QuarTeT software³ was used to search for telomeres and centromeres of two haplotype genomes.

Telomeres were detected by searching for the ‘AAACCCT’ unified repeat monomer of each terminal chromosome sequence. The approximate centromeric regions of both haplotypes were detected via the density of repeats and the Hi-C interaction heatmap.

Genome annotation and quality evaluation

Structural and functional predictions of coding genes in the assembled *R. persica* genome were performed using homology annotation, *de novo* annotation, and transcriptome-assisted annotation methods. Blast was used to compare the related species with the genome of *R. persica* with the E value (1e-5). Genewise v. 2.4.1 (<https://www.ebi.ac.uk/seqdb/confluence/display/THD/GeneWise>) was used to identify coding sequences and introns. Augustus⁴ (v. 3.3) was used for *de novo* gene annotation. TopHat v. 2.1.1 (<http://ccb.jhu.edu/software/tophat/index.shtml>) was used to align the RNA-seq to the reference genome. Finally, the gene set was generated by EVidenceModeler v. 1.1.1 (<http://evidencemodeler.github.io/>) and evaluated by BUSCO and other protein databases (i.e., SwissProt, KEGG, eggNOG, GO, and NR). TE sequences were searched using the software packages RepeatMasker (<http://www.repeatmasker.org>) and RepeatProteinMask. RepeatModeler (<http://www.repeatmasker.org>) and LTR-FINDER⁵ were used to construct the repeat sequence database. TRF⁶ was used to identify tandem repeat sequences.

Comparative genomic analyses

Whole genome duplication (WGD) events were estimated based on transversion of four-fold degenerate site and Ka/Ks values. NGenomeSyn software (<https://github.com/hewm2008> NGenomeSyn) was used for synteny analysis between *R. persica* and *R. chinensis* ‘Old Blush’ genomes. A total of 11 Rosaceae species including *R. persica* (Hap 1) was used to construct orthogroups (Extended Table 10) using Orthofinder (v 2.5.4)⁷. Homologous genes and gene families were identified using diamond. The 707 single-copy gene set of 11 Rosaceae species was also identified using Orthofinder.

References

- 1 Tang, H. B. *et al.* JCVI: A versatile toolkit for comparative genomics analysis. *Imeta*, doi:10.1002/imt2.211 (2024).
- 2 Yu, C. *et al.* Karyotype Analysis of Wild Species in Xinjiang, Northwestern China. *Journal of the American Society for Horticultural Science* **139**, 39-47, doi:10.21273/Jashs.139.1.39 (2014).

- 3 Lin, Y. Z. *et al.* quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Horticulture Research* **10**, doi:10.1093/hr/uhad127 (2023).
- 4 Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**, W435-439, doi:10.1093/nar/gkl200 (2006).
- 5 Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* **35**, W265-268, doi:10.1093/nar/gkm286 (2007).
- 6 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-580, doi:10.1093/nar/27.2.573 (1999).
- 7 Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238, doi:10.1186/s13059-019-1832-y (2019).