WAGENINGEN
UNIVERSITY & RESEARCH

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

Nature Plants

Cheng, Bixuan; Zhao, Kai; Zhou, Meichun; Bourke, Peter M.; Zhou, Lijun et al

https://doi.org/10.1038/s41477-025-01955-5

This publication is made publicly available in the institutional repository of Wageningen University and Research, under the terms of article 25fa of the Dutch Copyright Act, also known as the Amendment Taverne.

Article 25fa states that the author of a short scientific work funded either wholly or partially by Dutch public funds is entitled to make that work publicly available for no consideration following a reasonable period of time after the work was first published, provided that clear reference is made to the source of the first publication of the work.

This publication is distributed using the principles as determined in the Association of Universities in the Netherlands (VSNU) 'Article 25fa implementation' project. According to these principles research outputs of researchers employed by Dutch Universities that comply with the legal requirements of Article 25fa of the Dutch Copyright Act are distributed online and free of cost or other barriers in institutional repositories. Research outputs are distributed six months after their first online publication in the original published version and with proper attribution to the source of the original publication.

You are permitted to download and use the publication for personal purposes. All rights remain with the author(s) and / or copyright owner(s) of this work. Any use of the publication or parts of it other than authorised under article 25fa of the Dutch Copyright act is prohibited. Wageningen University & Research and the author(s) of this publication shall not be held responsible or liable for any damages resulting from your (re)use of this publication.

For questions regarding the public availability of this publication please contact
openaccess.library@wur.nl

nature plants

Article

https://doi.org/10.1038/s41477-025-01955-5

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

Received: 15 December 2023

Accepted: 26 February 2025

Published online: 4 April 2025

Check for updates

Bixuan Cheng[1], Kai Zhao[2], Meichun Zhou[1], Peter M. Bourke[3], Lijun Zhou[1], Sihui Wu[1], Yanlin Sun[1], Lifang Geng[1], Wenting Du[1], Chenyang Yang[1], Juntong Chen[4], Runhuan Huang[1], Xiaoling Tian[1], Lei Zhang[1], He Huang[1], Yu Han[1], Huitang Pan[1], Qixiang Zhang[1], Le Luo[1] & Chao Yu[1]

The cultivation and domestication of roses reflects cultural exchanges and shifts in aesthetics that have resulted in today's most popular ornamental plant group. However, the narrow genetic foundation of cultivated roses limits their further improvement. Wild *Rosa* species harbour vast genetic diversity, yet their utilization is impeded by taxonomic confusion. Here we generated a phased and gap-free reference genome of *Rosa persica* for phylogenetic and population genomic analyses of a large collection of *Rosa* samples. The robust nuclear and plastid phylogenies support most of the morphology-based traditional taxonomy of *Rosa*. Population genomic analyses disclosed potential genetic exchanges among sections, indicating the northwest and southwest of China as two independent centres of diversity for *Rosa*. Analyses of domestication traits provide insights into selection processes related to flower colour, fragrance, double flower and resistance. This study provides a comprehensive understanding of rose domestication and lays a solid foundation for future re-domestication and innovative breeding efforts using wild resources.

The domestication of ornamental plants by humans spans a history of approximately 5,000 years[1]. This process has not only created colourful civilizations but also witnessed the continuous pursuit of aesthetic values by human society. Successive rounds of breeding and selection have shaped the appealing traits of ornamental plants. Despite their appeal, a potential consequence of this process is a rapid decline in genetic diversity, a result of short-term selection pressure within a limited genetic pool[2], particularly associated with the reduction or loss of resistance. Wild related species play central roles in modern plant breeding[3]. Their rich genetic diversity can accelerate the improvement and deployment of new varieties, while introducing favourable alleles for targeted traits[4]. The strategy of introducing wild resources and expanding genetic backgrounds for the cultivation of ornamental plants has received considerable attention from breeders aiming to promote the sustainability of the ornamental plant industry[5].

Roses—known as the 'queen of flowers'—belong to the genus *Rosa* in the Rosaceae family. As the most successfully domesticated ancient ornamental plants, roses have been regarded as a symbol

[1]State Key Laboratory of Efficient Production of Forest Resources, Beijing Key Laboratory of Ornamental Plants Germplasm Innovation and Molecular Breeding, National Engineering Research Center for Floriculture and School of Landscape Architecture, Beijing Forestry University, Beijing, China. [2]College of Life Sciences, Fujian Normal University, Fuzhou, China. [3]Plant Breeding, Wageningen University & Research, Wageningen, The Netherlands. [4]Key Laboratory of Plant Diversity and Specialty Crops, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, China.
e-mail: yuchao@bjfu.edu.cn

of love and beauty from antiquity to modern times[6]. Today, roses are the most extensively cultivated ornamental plants. Their annual global trade represents 30% of the cut flower market (www.aiph.org/statistical-yearbook), and the number of rose cultivars has exceeded 35,000 and continues to grow. Nevertheless, the origin of most rose cultivars dates back to relatively recent times[7,8], following the breeding renaissance, which occurred in the late eighteenth century. Crosses among ancient Chinese roses, wild *Rosa* species and European old cultivars resulted in the first notable revolution in rose breeding, introducing traits of recurrent blooming as well as rich colour and fragrance. Until today, with only eight to ten wild *Rosa* species contributing to the genesis of modern roses[9,10], the cultivated rose population has revealed a constrained genetic foundation[11–13]. This makes it difficult to achieve breakthroughs in breeding efforts.

Global climate changes are driving rose breeding practice toward strong resistance and low maintenance, which urgently necessitates the introduction of genetic resources from wild relatives. The utilization of wild resources is based on comprehensive understanding of both their phylogenetic relationships and evolutionary history. However, wild *Rosa* species have not been systematically studied at the genus level in sufficient detail. Abundant phenotypic variation, together with frequent hybridization and polyploidization events at the genus level, has led to widespread discrepancies and taxonomic confusion in the genus *Rosa*[14–16]. The current conventional taxonomy of the genus *Rosa* still relies on work based on morphological characters, categorizing this genus into 2–4 subgenera, and subgenus (subg.) *Rosa* is further divided into 10–12 sections[17–19]. Numerous studies have attempted to clarify the phylogenetic relationships within this genus at the molecular level, but these efforts have consistently resulted in conflicting conclusions, failing to support the current morphologically based taxonomy. This has hindered the further utilization of wild *Rosa* resources.

With existing studies mostly focusing on rose cultivars, the assemblies of several *Rosa* genomes have traced the genetic origin of modern roses[10,20], thus laying the foundation for understanding the molecular mechanisms underlying important traits[21–24]. By contrast, most wild *Rosa* resources remain understudied, yet they represent invaluable materials for the breeding and phylogenetic research of this genus[16]. China is one of the primary centres of distribution and diversity of *Rosa* species, with more than 95 locally distributed species[25]. The development of sequencing technology has made it possible to study these wild resources at the genus level. In this study, we used sequence-based whole-genome analyses on a large collection of *Rosa* samples. Based on a telomere-to-telomere assembly of the early diverging wild *Rosa* species *Rosa persica*[26], we propose a robust and comprehensive *Rosa* phylogeny. By uncovering the botanical origin, phylogenetic relationships and evolutionary history of *Rosa*, this study lays a solid foundation for the utilization of wild *Rosa* resources. The aim is to assist in the re-domestication and revolutionary breeding works of modern roses.

Results

Telomere-to-telomere genome assembly of *R. persica*

To better explore the systematic evolution of *Rosa* and the domestication process of important traits, we generated a phased and gap-free genome of *R. persica* ($2n = 2x = 14$). Flow cytometry analysis revealed that *R. persica* had an estimated genome size of 0.37 Gb (Supplementary Table 1 and Supplementary Fig. 1a), which was similar to the size estimated by K-mer analysis (0.35 Gb, Supplementary Fig. 1b). We obtained a total of 86.56 Gb (247×) PacBio high-fidelity (HiFi) reads and 55.37 Gb chromosome conformation capture (Hi-C) data (Supplementary Table 2) for genome assembly. Initial assembly yielded two different haplotypes for the diploid *R. persica* genome (Supplementary Table 3 and Supplementary Fig. 2). A haplotype-resolved assembly was obtained with zero gaps in the genome after scaffolding contigs using Hi-C data and gap-closing with quarT2T[27] (Supplementary Table 4 and Supplementary Fig. 3). Final assembly obtained total lengths of 364.44 Mb for haplotype 1 (Hap 1) and 363.41 Mb for haplotype 2 (Hap 2). Using the unified telomere repeat ('AAACCCT') as sequence query, we identified 28 telomeres (14 for each haplotype; Supplementary Fig. 4). The 14 centromeric regions of both haplotypes were also detected, ranging from 1.1 to 4.3 Mb, with an average length of 2.2 Mb (Supplementary Table 5).

The quality and completeness of the assembly were evaluated (Supplementary Table 6) and compared to published *Rosa* genomes (Supplementary Table 9). Evaluation of Benchmarking Universal Single-Copy Orthologs (BUSCO) showed that 98.90% of the orthologous gene sets were present for both haplotype genomes. Second, the quality values were 66.16 (Hap 1) and 65.42 (Hap 2), which highlighted the high accuracy of the assembly. Third, the long terminal repeat (LTR) assembly index (LAI) and mapping rates of both haplotypes indicated that the assembly was of very high quality (Supplementary Table 6).

Totals of 32,351 (Hap 1) and 32,408 (Hap 2) protein-coding genes were predicted (Supplementary Table 7). The *R. persica* genome annotated 52–54% of repetitive sequences, most of which are LTR sequences (LTRs) (~34%) (Supplementary Table 8). LTR contents are the major factors contributing to variations in genome sizes[28,29]. Comparison of published *Rosa* genomes shows that repeat sequences contribute the majority of *Rosa* genomes (47.92–67.90%; Supplementary Table 9). Further comparison between *R. persica* and *R. chinensis* 'Old Blush' showed that *R. chinensis* 'Old Blush' had higher LTR contents and experienced a large-scale LTR insertion burst in recent times (Supplementary Fig. 5a), which may have led to its larger genome size.

R. persica did not undergo a whole genome duplication event (Supplementary Fig. 5b), which is consistent with the results of collinearity analysis of *R. persica* and *R. chinensis* 'Old Blush'[10] (Supplementary Fig. 5c). Further analysis using 11 Rosaceae species identified a total of 11,072 shared orthogroups, 707 of which were entirely single-copy genes (Supplementary Table 10). This core Rosaceae single-copy gene set was used for further phylogenetic studies.

Resequencing and phylogenetic analyses of *Rosa* accessions

To fully examine the phylogenetic relationships within this genus, this paper presents a large collection of *Rosa* samples (Fig. 1). Because of the similarity of phenotypic characteristics, misidentifications in *Rosa* accessions have been common. First, we conducted a detailed examination of *Rosa* species recorded in *Flora of China*[25]. Based on years of field investigations, combined with phenotypic measurements, specimen comparison and phylogenetic results, we have revised the classification of several species, as detailed in Supplementary Notes. Subsequent analyses were based on current revisions, which were also supported by previous studies (Supplementary Notes). After meticulous identification, we assembled a total of 205 *Rosa* accessions, covering all 3 subgenera and 10 sections within the genus *Rosa* (Extended Data Figs. 1 and 2). This collection encompasses 84% (80/95) of *Rosa* species documented in *Flora of China* (Supplementary Table 11). In addition, this study includes newly published *Rosa* species/varieties[30–33], as well as subspecies, varieties and cultivars that were recorded in local flora literature but not included in the *Flora of China*. We also included 47 rose cultivars, 27 of which are ancient roses, and examples of others are 'Eyes for you' and 'Tianshan' cultivars.

We compiled ploidy information for most of our accessions (Supplementary Table 12), among which the ploidy levels of 108 *Rosa* species were confirmed through chromosome observation. The ploidy information of other materials was accessed through previously reported studies. The ploidy of collected materials ranged between $2x$ and $10x$. However, most were diploid (~70%).

Whole-genome resequencing was performed on the above materials, and high-quality Illumina sequencing data with an average read-depth of 30× were obtained. In addition, we collected data of 10 published accessions, resulting in a total of 215 high-quality datasets for the genus *Rosa* (Supplementary Table 13). These data were aligned

https://doi.org/10.1038/s41477-025-01955-5

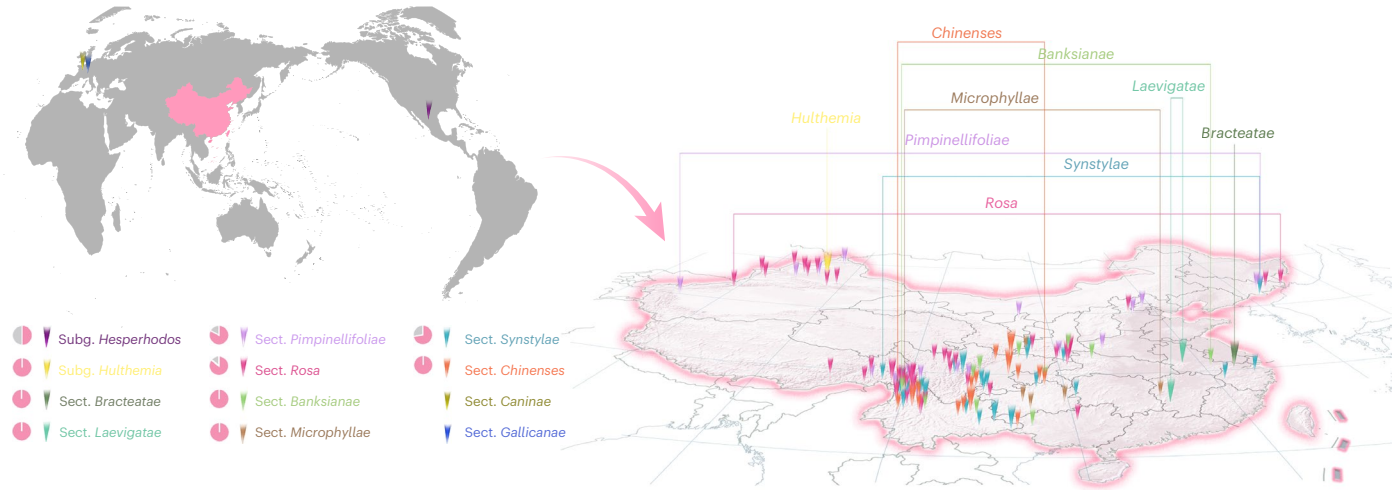


Fig. 1 | Geographical distribution of core *Rosa* accessions. The collected subg. *Hesperhodos* accession originates from North America. Accessions in sect. *Caninae* and sect. *Gallicanae* originate from Europe, as shown in the left plate. All other accessions were collected from China, as shown in the right plate. The small pie charts correspond to the percentages of collected *Rosa* accessions covered in the subgenus/section. Map data from the National Standard Map Service System (http://bzdt.ch.mnr.gov.cn/; source number GS(2016)1613) (left) and Google Earth, SIO, NOAA, US Navy, NGA, GEBCO (right).

to the reference genome of *R. persica* Hap 1 for variant calling, and high-quality single-nucleotide polymorphisms (SNPs) were obtained for subsequent analysis.

Based on the genomic location of 707 single-copy nuclear genes, we obtained 6,048 conserved SNPs (single-copy SNPs), which were used to construct a robust phylogenetic tree across the whole genus (Fig. 2 and Supplementary Fig. 6). This phylogeny assigns a distinct basal position to subg. *Hulthemia* (clade A). The largest subgenus (subg. *Rosa*) was resolved as paraphyletic with two well-supported clades (clades B and C). Clade B consisted of *R. stellata* from subg. *Hesperhodos* and a clade of section (sect.) *Pimpinellifoliae*, which is mainly distributed in northern China. These diverging clades appeared as early derived groups within subg. *Rosa*. Species from other sections formed clade C. Specifically, sect. *Bracteatae*, sect. *Laevigatae* and sect. *Banksianae*, all mainly distributed in southern China, appeared as sister groups to each other and formed a clade C1. Taxa from sect. *Microphyllae* belonged to the diverging clade (clade C2), while clade C3 included all taxa from sect. *Rosa*. Clade C4 consisted of a wide range of recently derived materials, including a subclade of 'Tianshan' cultivars with ancestry from *R. laxa*; a subclade consisting of accessions of European background, including sect. *Caninae*, sect. *Gallicanae* and *R. arvensis*; and finally, the nested subclade of sect. *Synstylae*, sect. *Chinenses* and cultivated roses.

We also assembled the plastomes of 205 *Rosa* accessions based on sequencing data and used annotated coding sequences to reconstruct a plastid phylogeny of *Rosa*. Phylogenies inferred from plastid data produced similar results to nuclear-based analyses, where samples from different sections were basically clustered together (Extended Data Fig. 3). However, incongruencies were still evident between trees, and the positions of certain material were rearranged. Examples are *R. spinosissima* and *R. platyacantha* of sect. *Pimpinellifoliae*, *R. pseudobanksiae* and *R.* × *fortuneana* of sect. *Banksianae*, and *R. glomerata* of sect. *Synstylae*. Among the incongruencies between the nuclear and plastid trees, subg. *Hesperhodos* (*R. stellata*) was identified as the basal position of the plastid tree, while it was in the second-diverging clade of the nuclear tree. Subg. *Rosa* was paraphyletic in the plastid tree but monophyletic in the nuclear tree. Subg. *Hulthemia* (*R. persica*) was included within sect. *Pimpinellifoliae* in the chloroplast tree, while it was a sister to what remained of the genus in the nuclear tree. Two major clades were found, and 'Tianshan' cultivars were clustered within a well-resolved subclade, together with their parental material *R. laxa*. Sect. *Synstylae*, sect. *Chinenses* and cultivated roses together formed a nested subclade.

Ancestral state reconstruction of major traits in *Rosa*

The reconstruction of ancestral character states for five morphological characters provide insights into the evolutionary history of important traits of *Rosa*. Using maximum parsimony and maximum likelihood methods, we obtained similar results, namely, that solitary, simple flowers and yellow colour with no petal spots are inferred as ancestral traits of *Rosa*. The leaflet number underwent multiple transition events in the phylogeny, and the highest probability of the ancestral state was seven leaflets ($P = 0.37$) (Fig. 3a). Regarding the number of petals, both methods suggested that the ancestral state was simple flower (Fig. 3b). The double flower trait was retained in the common ancestor node of sect. *Synstylae*, sect. *Chinenses*, sect. *Caninae* and cultivated roses. The ancestral trait for the flower inflorescence organization was solitary flowers (Fig. 3c). The transition from solitary flowers to corymb or compound inflorescence may have occurred twice in the phylogeny. The first transition was observed on the branch leading to sect. *Banksianae*, while the second was in the clades of sect. *Synstylae* and cultivated roses. Potential genetic exchanges were observed between sect. *Banksiae* and sect. *Synstylae* (Fig. 5b). It is thus possible that this trait was first acquired from sect. *Banksiae* and subsequently passed down to these recently derived groups. Although *R. persica* is relatively primitive within *Rosa*, both methods match in that they suggest that the ancestral state had no petal spot. The petal spot for wild taxa appeared only in *R. persica* and was introduced into cultivated roses, particularly in the 'Eyes for you' cultivar group, through recent breeding process (Supplementary Fig. 7).

For flower colour, to reconstruct ancestral states, we used grading traits as discrete characters and RGB values (red, green and blue color model) as continuous characters (Fig. 3d). The results consistently identified yellow flowers as the ancestral trait of *Rosa*. Flower colour transitioned from yellow (subg. *Hulthemia* and subg. *Pimpinellifoliae*) and white (sect. *Banksianae*) to red/pink colours during evolution and domestication processes. During domestication and selection, traits such as double flowers and vibrant colours have been long-term breeding targets, which potentially lead to considerable differences in the characteristics of modern roses compared to their ancestral states. This reflects the influence of human selection on the evolution of ornamental traits. It also indicates that the contributions of primitive materials, such as *R. persica*, to modern roses are relatively small.

Population structure discloses differentiation within *Rosa*

The results of a principal component analysis (PCA) based on sequencing data of 215 samples indicated that PC1 (51.1%) generally separates

https://doi.org/10.1038/s41477-025-01955-5

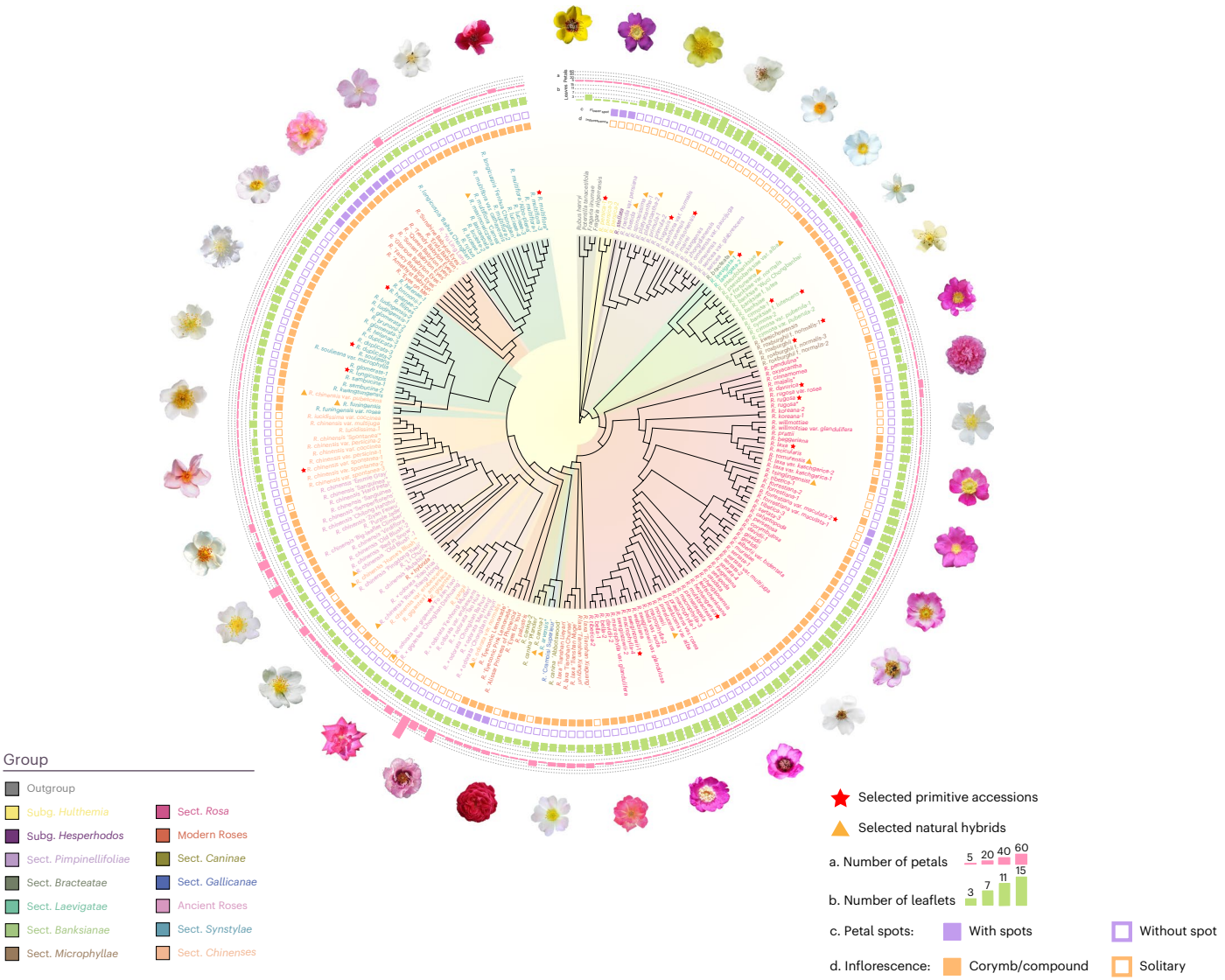


Fig. 2 | Phylogenetic relationships of 215 *Rosa* accessions based on 6,048 SNPs of 707 single-copy nuclear genes. The topology was constructed using a maximum likelihood method with 1,000 bootstrap replicates to assess the support of branches. Background colours represent accessions from different botanical groups. We selected core accessions based on genetic analyses and morphological traits, which are marked with stars and triangles following their scientific names. The outer circle represents the floral traits of selected accessions.

early differentiated accessions from recently differentiated accessions (Fig. 4a). The early diverging accessions from subg. *Hulthemia*, subg. *Hesperhodos*, sect. *Pimpinellifoliae* and sect. *Rosa* clustered together, while the recently differentiated sect. *Synstylae*, sect. *Chinenses* and cultivated roses also clustered together.

Population structure analysis was conducted using ADMIXTURE (v.1.3.0). With the optimal solution at nine populations ($K = 9$) (Supplementary Table 14), the population structure corresponded well with the phylogenetic tree (Fig. 4d). In addition, the topological differences between the chloroplast and nuclear tree were also supported by the results of population structure analysis. Species that differentiated early in the phylogenetic tree showed a relatively consistent genetic background, indicating fewer inter-specific hybridization events. *R. platyacantha*, *R. foetida* and *R. spinosissima* shared a hybrid background of both sect. *Pimpinellifoliae* and sect. *Rosa*, suggesting that these accessions may be natural inter-specific hybrids. Potential genetic exchanges were also observed between sect. *Banksianae* and sect. *Synstylae*, with *R. pseudobanksiae*, *R.* × *fortuneana* and *R. uniflora* showing mixed genetic compositions. Accessions with European background in sect. *Caninae* and sect. *Gallicanae* had hybrid origins. The genetic composition of cultivated roses almost entirely originated from sect. *Chinenses* and sect. *Synstylae*, further corroborating that species in these two sections were the main contributors to the cultivated roses[34].

We further divided the 215 accessions into an early differentiated group (group 1) and a recently differentiated group (group 2), based on the results of population structure, PCA and phylogenetic analysis. Moreover, within group 2, the wild species (group 2-W) and cultivated roses (group 2-C) were separated. Linkage disequilibrium analysis of the three groups indicated that group 1 showed the fastest linkage disequilibrium decay, followed by group 2-W, with group 2-C showing the slowest linkage disequilibrium decay (Fig. 4b). Calculation of nucleotide diversity (π) for the three groups showed that group 1 had the highest nucleotide diversity ($\pi = 1.33 \times 10^{-2}$), followed by group 2-W ($\pi = 1.13 \times 10^{-2}$). With a value of 1.08×10^{-2}, cultivated roses showed the lowest nucleotide diversity (Fig. 4c). These results suggest that artificial selection during the domestication process of roses has led to a decrease in population genetic diversity.

https://doi.org/10.1038/s41477-025-01955-5

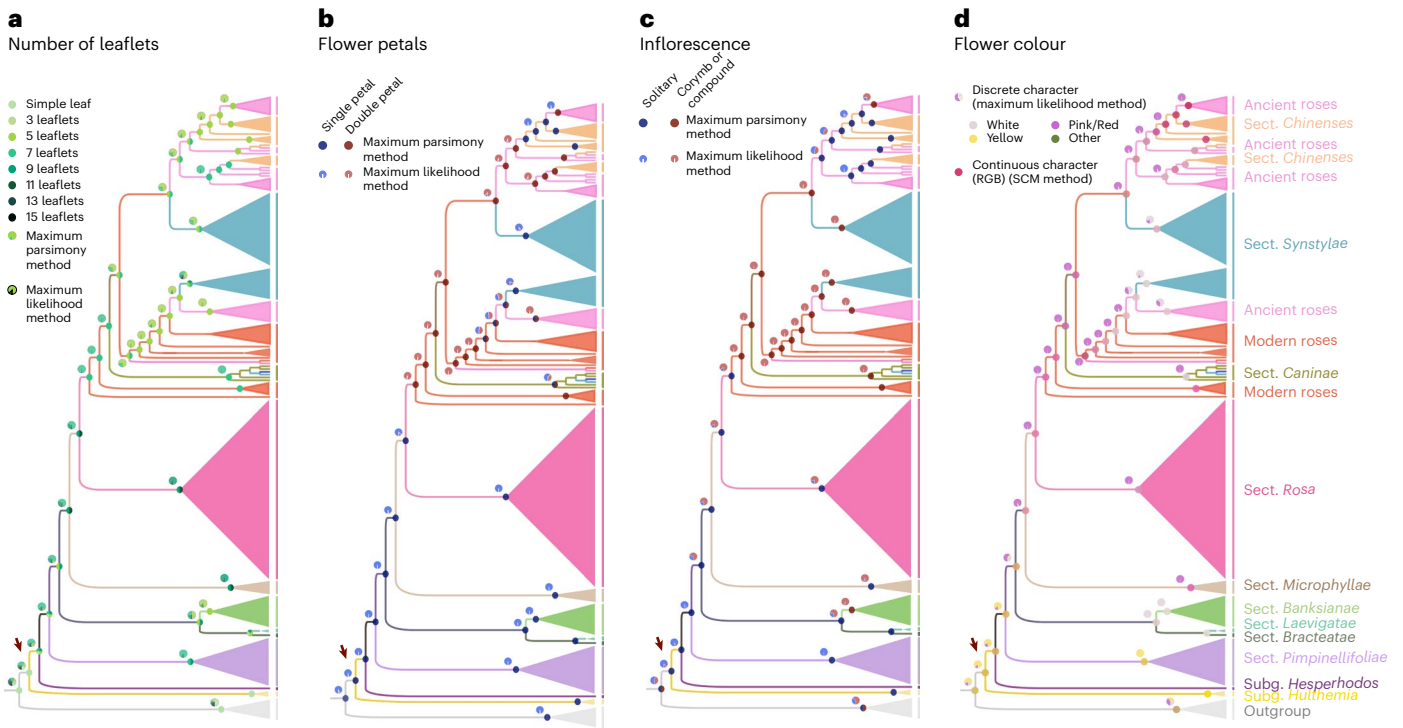


Fig. 3 | Ancestral state reconstruction of the four selected traits of *Rosa*, performed using both the maximum parsimony method (at the node of the tree) and the maximum likelihood method (above the node of the tree) for the backbone of the single-copy SNP tree. a–d, The colour keys and pie diagrams 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 the number of leaflets (**a**), flower petals (**b**), inflorescence type (**c**) and flower colour (**d**). Both discrete characters (at the node of the tree) and continuous characters (RGB values, above the node of the tree) were used to reconstruct the ancestral state of flower colour traits. The ancestral state reconstruction of petal spots is presented in Supplementary Fig. 7. SCM, stochastic character mapping.

Retracing the origin and demographic history of *Rosa*

Geographical information of *Rosa* species indicated two distinct distribution areas in China (Fig. 5e and Extended Data Fig. 4). TreeMix (v.1.13) and structure analyses found no gene flow among early-derived species distributed in the northwest (subg. *Hulthemia* and sect. *Pimpinellifoliae*) and southwest (sect. *Banksianae*) of China; the fixation index (F_{ST}) value of 0.16 was found between sect. *Pimpinellifoliae* and sect. *Banksianae* (Fig. 5c). We thus suggest that there are two diversification centres for the genus *Rosa*, one in the northwest and the other in the southwest of China. TreeMix and *D*-statistic (Fig. 5a,b and Supplementary Table 15) identified gene flow between sect. *Rosa* and sect. *Pimpinellifoliae*, both of which are mainly distributed in the northern area of China. Gene flow between sect. *Synstylae* and sect. *Banksianae* was also detected, with a low fixation index (F_{ST} = 0.11). According to the similar geographical distributions between these sections, it can be inferred that sect. *Rosa* originated from sect. *Pimpinellifoliae*, while sect. *Synstylae* originated from sect. *Banksianae*. Little genetic differentiation was found between sect. *Chinenses* and sect. *Synstylae* (F_{ST} < 0.05). Together with the phylogenetic results, it is suggested that sect. *Chinenses* originated from sect. *Synstylae* and that the divergence between them happened relatively recently.

We selected sections with sufficient sample size and performed effective population size analysis using Stairway Plot (v.2.1) (Fig. 5d). Previous studies have shown that *Rosa* species diverged at approximately 6 Ma (million years ago)[22,23,35]. Population dynamics analysis showed that the first population bottleneck in *Rosa* occurred after approximately 6 Ma, during the transition period between the Miocene and the Pliocene[36]. This population bottleneck promoted differentiation among *Rosa* sections.

During this period, sect. *Banksianae* underwent significant differentiation from others. This differentiation may have been caused by significant geographic differences between the north and south of China during the late Tertiary Himalayan orogeny (approximately 23–1.6 Ma), thus accelerating population divergence. The second population bottleneck occurred during the Last Glacial Maximum (approximately 26.5–19 ka (thousand years ago)), which was characterized by a significant decrease in global temperature and precipitation[37], resulting in a continuous decline in *Rosa* population size. The effective population size of sect. *Banksianae* recovered at around 0.2 Ma and remained stable thereafter, possibly because of its wide geographic distribution and greater number of species within the group. Sect. *Synstylae*, sect. *Chinenses* and rose cultivars in group 2 showed similar trends, indicating that these populations did not undergo significant differentiation.

Identification of genomic footprints of domesticated traits

To further examine selected loci during the rose domestication processes, we conducted selective sweep analysis between group 1 and group 2, as well as between group 2-W and group 2-C (Fig. 6b and Supplementary Table 16). Selective sweeps that were identified between group 1 and group 2 were considered to be involved in the evolutionary history and domestication of *Rosa* species, and genes enriched for the Gene Ontology categories 'trichome branching', 'toxin catabolic process' and 'glutathione transferase activity' (Supplementary Fig. 8a). Sweeps identified between group 2-W and group 2-C were primarily related to the recent domestication of cultivated roses. Genes under selection between group 2-W and group 2-C were enriched in 'response to fructose', 'stem cell division' and scent-related pathways, such as 'sesquiterpene metabolic process' and 'terpene synthase activity' (Supplementary Fig. 8b).

Flower colour is a critical domestication trait that has undergone a transformation from the ancestral traits of white/yellow to red and pink

https://doi.org/10.1038/s41477-025-01955-5

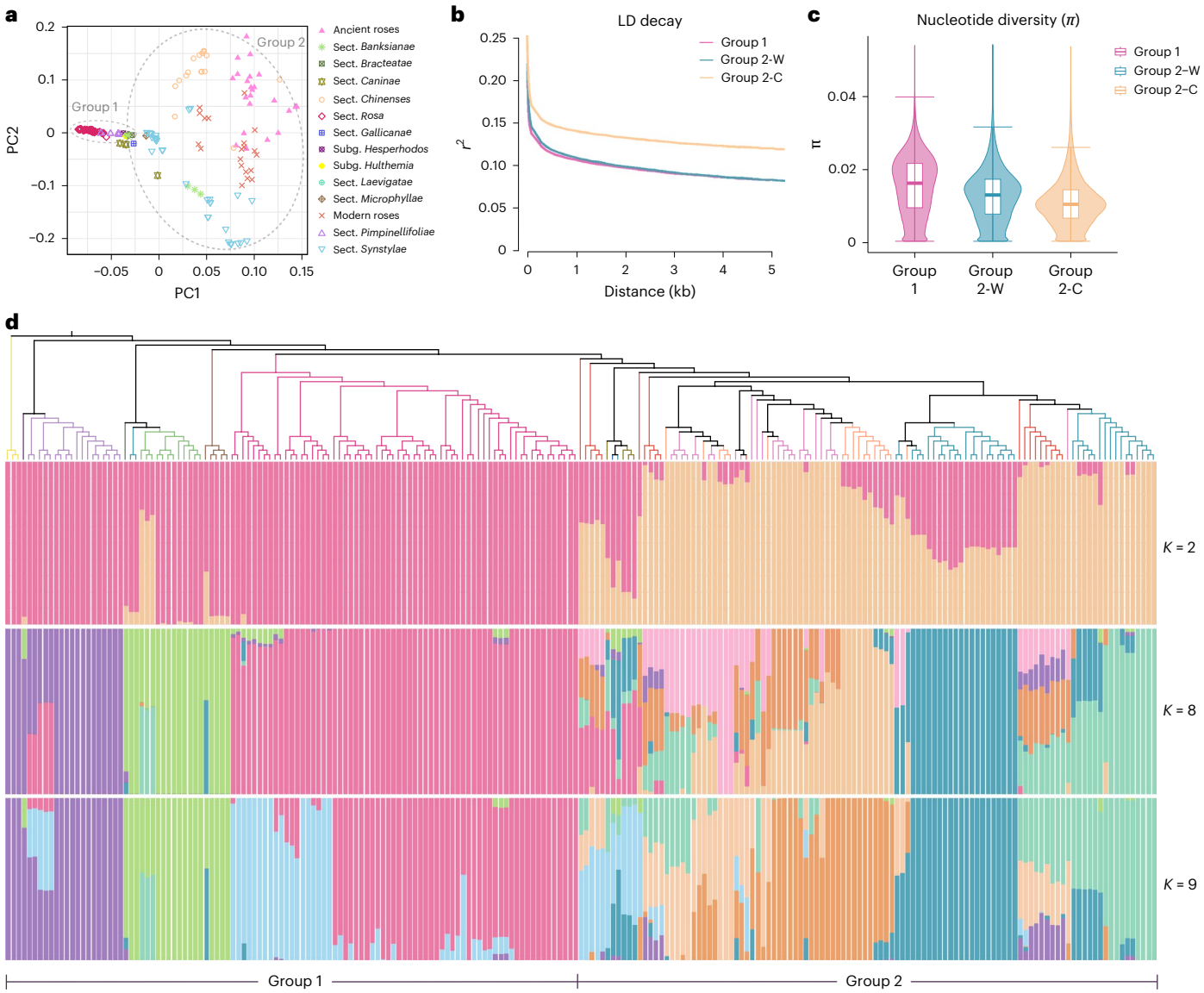


Fig. 4 | Population structure of *Rosa* accessions. a. PCA of 215 *Rosa* accessions. PC1 and PC2 account for 51.1% and 18.8% of the total variation, respectively. **b**. Linkage disequilibrium (LD) decay pattern of different groups. **c**. Violin plots of nucleotide diversity for different groups (centre line, median; box limits, upper and lower quartiles). **d**. Population structure of 215 *Rosa* accessions, corresponding with the phylogenetic tree. The bar plots represent the percentage of membership (*q*) for each group identified at $K = 2$, $K = 8$ and $K = 9$.

colours that predominate in cultivated roses today. The formation of the petal colour is related to both carotenoid and flavonoid pathways. Here, genes involved in carotenoid synthesis were selected, including *15-cis-phytoene desaturase*, *phytoene synthase* and *zeta-carotene isomerase*. Genes such as *flavonol synthase*, *flavanone 3′-hydroxylase* and *anthocyanidin 3-O-glucosyltransferase* were identified in the selected regions on Chr1, Chr2 and Chr4. These genes are involved in flavonoid biosynthesis and glycosylation pathways[38,39] (Supplementary Table 17).

Many V-myb avian myeloblastosis viral oncogene homolog (MYB) transcriptional factors in the biosynthesis of anthocyanin and proanthocyanidins have been functionally characterized in roses[10,21,40–42]. We also found several R2R3-MYBs under selection. Phylogenetic analysis disclosed that these genes were clustered into S5, S7, S20 and S21 subfamilies (Supplementary Fig. 9). Haplotypes of two identified MYB genes, *AO02073.1* (group 1 versus group 2) and *AO14329.1* (group 2-W versus group 2-C), were significantly correlated with petal colours (Supplementary Figs. 10 and 11). Most of the red/pink petal colours were associated with Hap 1 of both genes, while petals with white/yellow colours were associated with other haplotypes. It is worth noting that these haplotypes were specific to certain groups of materials (Supplementary Figs. 10b and 11b), indicating their important roles in floral colour differentiation during the evolutionary process of roses. Specifically, Hap 1 of both genes represented a unique genetic feature of wild roses in group 1 because they were exclusive in materials from this group.

The fragrance of modern roses has been inherited from multiple ancestor species historically involved in rose breeding in the nineteenth century. However, many characteristic fragrances, especially in cultivars for cut flower productions, were lost during the breeding process in the twentieth century because the main breeding goals shifted toward long vase life and disease resistance. Here, in addition to genes responsible for flower colour, we also found genes associated with biosynthetic pathways of volatile compounds in selective sweeps. For instance, a *terpenoid synthase gene* (*TPS*) on Chr7, which was reported to be associated with the sesquiterpene germacrene D[10,43],

https://doi.org/10.1038/s41477-025-01955-5

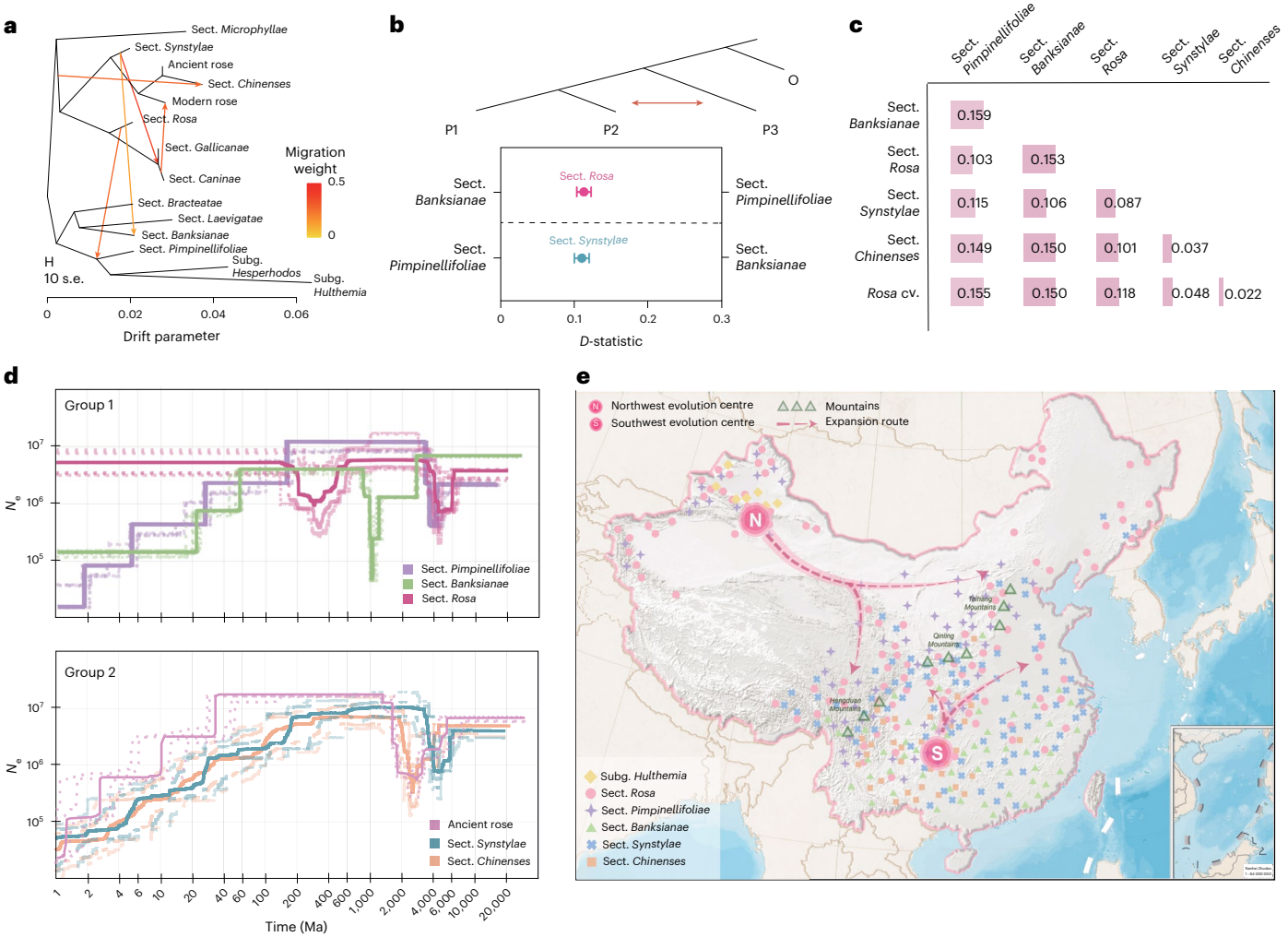


Fig. 5 | Demographic history of *Rosa*. **a**,**b**, Introgression among different *Rosa* populations using TreeMix (**a**) and *D*-statistic (**b**). **c**, Fixation index (F_{ST}) between different *Rosa* populations. The numbers in squares represent F_{ST} values. **d**, Effective population size for *Rosa* populations estimated using stairway plots. Solid lines indicate the median of 200 inferences. Dashed lines indicate 95% and 75% confidence intervals of the inference. **e**, Postulated evolution centre and expansion routes of *Rosa* in China. Geographical distribution of *Rosa* accessions under six botanical groups is shown in different coloured shapes. Map in **e** from the National Standard Map Service System (http://bzdt.ch.mnr.gov.cn/; source number GS(2016)1613).

was under selection. In the selective sweep region between group 2-W and group 2-C, genes related to floral scent were all distributed on Chr5, and loci overlapped with the previously identified quantitative trait loci (QTLs) of 2-phenyl ethanol[44]. We identified the key enzyme gene *phenylacetaldehyde reductase*, which is involved in the biosynthesis of 2-phenyl ethanol and rhodinol in roses[45–47]. Other genes related to terpenoid and phenypropanoid biosynthetic pathways, such as *nerolidol synthase*, *eugenol synthase*, and *caffeoyl-CoA O-methyltransferase*, were also under selection.

In contrast to wild accessions, in cultivated roses, the characteristic of continuous flowering is particularly valued. We identified genes related to the continuous flowering trait between group 2-W and group 2-C, indicating that the continuous flowering trait is the result of recent selection events, rather than of long-term natural selection[48]. The *Koushin* (*KSN*) gene on Chr3 (14.6–15.6 Mb), which has been functionally characterized as one of the key factors for the formation of continuous flowering[10,24,48,49], was under selection. Results based on QTL analysis indicated that the continuous flowering trait is controlled by at least two loci[10,50,51]. We also identified *flowering-promoting factor 1* and *flowering time control protein* at the selection loci on Chr4 (32–34 Mb). *Flowering-promoting factor 1* has been proven to promote flowering in *Arabidopsis*[52] and *Fragaria vesca*[53]. *Flowering time control protein* may be involved in regulating flowering through a thermosensory pathway[54]. This locus overlapped with the QTL region for flowering date identified previously[55].

Genes related to floral organ development were identified in selection sweep loci on multiple chromosomes. In the genus *Rosa*, these genes may be associated with domestication traits such as petal number and inflorescence. The results of a genome-wide association study (GWAS) disclosed key QTLs for petal number (Fig. 6a and Supplementary Fig. 12). *APETALA 2-like* was identified in the GWAS and F_{ST} overlapping locus on Chr3 (9.8–11.8 Mb) (Supplementary Table 18). The *APETALA 2-like* has been proven to be regulated by miR172 and determines the single/double flower trait in roses[10,24,56,57]. In this study, this gene was identified in both F_{ST} groups, indicating that it has been under continuous selection pressure. Interestingly, two *SQUAMOSA-promoter binding-like* (*SPL*) genes were also found to be under selection on Chr1, and *SPL8* (*AO01449.1*) was predicted to be the putative target of miR156[10].

We also focused on genes associated with disease resistance mechanisms and certain defence response-related genes within the selected intervals. Gene homologues encoding proteins containing nucleotide-binding sites and leucine-rich repeat sequences and proteins with a central nucleotide-binding domain (NB-ARC) have been found

https://doi.org/10.1038/s41477-025-01955-5

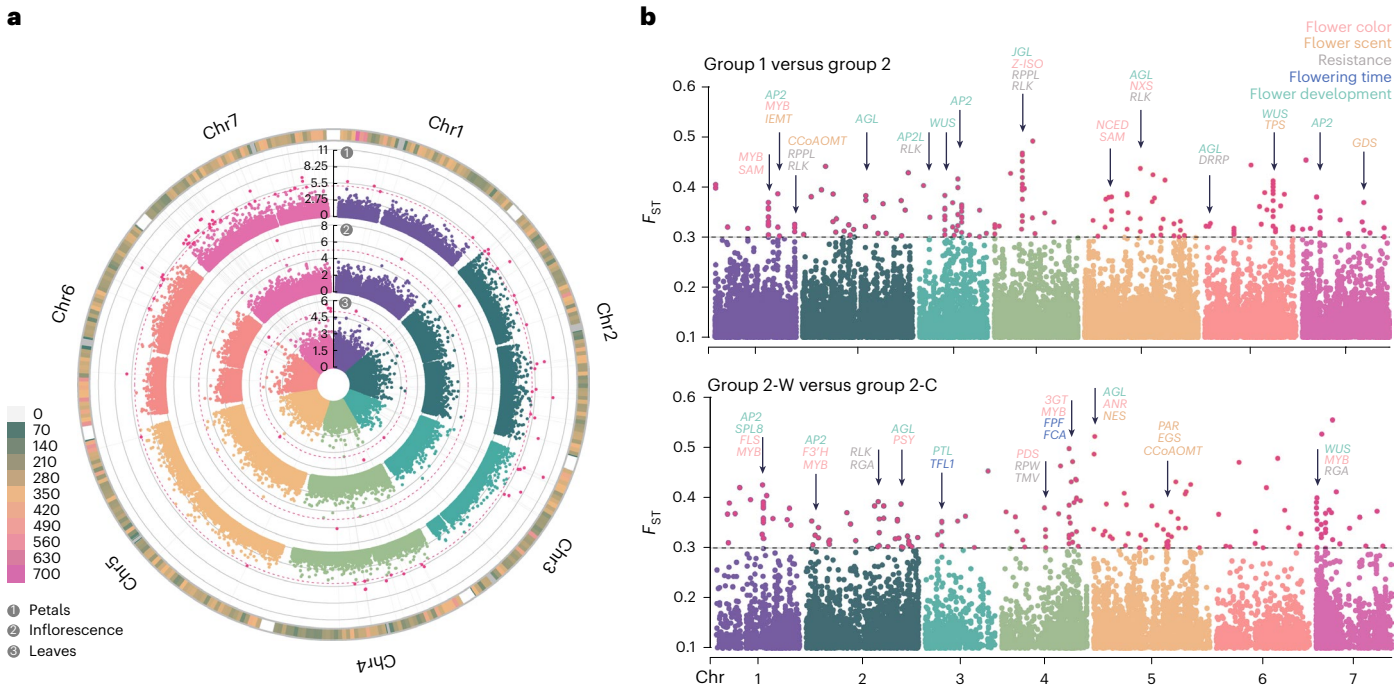


Fig. 6 | Genome-wide identification and selective sweep analysis across *Rosa* populations. a, GWAS on the number of petals (outer track), inflorescence types (middle track) and the number of leaflets (inner track). The red dotted line denotes the genome-wide significance threshold of 5.2, as determined using a Bonferroni test. **b**, Selective sweep analysis to identify candidate genes. The upper selective sweeps contrast group 1 and group 2 populations (improvement and domestication sweeps), while the lower panel contrasts group 2-W and group 2-C populations (domestication sweeps). Gene abbreviations: *SAM*, *S-adenosyl-L-methionine-dependent methyltransferase*; *AP2*, *APETALA2*; *CCoAOMT*, *Caffeoyl-CoA O-methyltransferase*; *RPPL*, *Disease resistance RPP-Like*; *RLK*, *RLK-Pelle-DLSV family*; *AGL*, *AGAMOUS-Like*; *AP2L*, *APETALA2-Like*; *WUS*, *WUSCHEL*; *JGL*, *JAGGED-Like*; *Z-ISO*, *15-cis-zeta-carotene isomerase*; *NCED*, *9-cis-epoxycarotenoid dioxygenase NXS*, *Neoxanthin synthase*; *DRRP*, *Disease resistance response protein*; *TPS*, *Terpenoid synthases gene*; *GDS*, *Germacrene D synthase*; *SPL*, *Squamosa promoter-binding-like protein*; *FLS*, *flavonol synthase*; *F3′H*, *flavanone 3′-hydroxylase*; *PSY*, *phytoene synthase*; *PTL*, *Petal Loss*; *TFL1*, *TERMINAL FLOWER 1*; *PDS*, *15-cis-phytoene desaturase*; *RPW*, *Powdery mildew resistance protein*; *TMV*, *TMV resistance protein*; *3GT*, *anthocyanidin 3-O-glucosyltransferase*; *FPF*, *Flowering-promoting factor*; *FCA*, *flowering time control protein*; *ANR*, *Anthocyanin regulatory protein*; *NES*, *nerolidol synthase*; *PAR*, *phenylacetaldehyde reductase*; *EGS*, *eugenol synthase*.

within several distinct peaks; all of these peaks are associated with general disease resistance mechanisms[58–60]. In the selective regions between group 2-W and group 2-C, the most disease-resistance-related proteins were found to be distributed on Chr4, and a Bet v1 gene cluster was found in addition to nucleotide-binding sites and leucine-rich repeat sequence genes and NB-ARC genes (Supplementary Table 17). Bet v1 belongs to the PR-10 disease-related protein family; its expression can be induced by pathogen infection, mechanical injury or abiotic stress[61,62].

Discussion

Towards a refined taxonomy of the genus *Rosa*

Linnaeus noted 'Mihi videtur naturam miscuisse plures vel lusu ex uno plures formasse' (It seems to me that nature has mixed several of them or, by game, formed several from one) when referring to the genus *Rosa* in *Species Plantarum*[63]. The identification and taxonomic analyses of *Rosa* species remain a true challenge, even today. Research has indicated that misidentifications of samples have resulted in discrepancies and low comparability between different phylogenetic studies of the genus *Rosa*[26]. In this study, after years of field investigations, a comprehensive collection of *Rosa* accessions was used, and these materials were re-identified and classified by two means: (1) based on specimens and morphological observations, materials with stable phenotypes from the *Flora of China* were merged and treated as the same species (Supplementary Notes), and (2) the wild types, varieties, and cultivars were finely divided. Population structure analysis further validated our accurate identification of the materials. For example, the single-petaled *R. odorata* and *R. chinensis* are wild materials with a homozygous genetic background, while the double-petaled *R. chinensis* 'Old Blush' originated from human selection and shows a stronger admixture component. In previous studies, these cultivars may have been treated as wild varieties, thus affecting the accuracy of the results.

Based on plant identification and whole-genome re-sequencing data, we present a comprehensive and robust phylogenetic analysis of the genus *Rosa*. Numerous phylogenetic analyses based on nuclear and plastid genes have suggested that the subg. *Rosa* was not monophyletic and that subg. *Hesperhodos*, subg. *Hulthemia* and subg. *Platyrhodon* should be classified at the sectional level[14,64,65]. Our phylogeny supports that most sections delimited by ref. 18 are monophyletic and that subg. *Hulthemia* shows a distinct basal position in the nuclear-based phylogeny, while subg. *Hesperhodos* formed an isolated clade at the base of the plastid phylogenetic tree. Recent studies have shown that subg. *Hesperhodos* and subg. *Hulthemia* have shown the highest genetic distance from other roses, suggesting that they are the only remaining descendants of an ancient group of roses[26,66]. Thus, we propose that these two subgenera should be retained.

Many phylogenetic studies found that sect. *Synstylae* and sect. *Chinenses* fall within the same clade[64,66–68]. Our results also showed that sect. *Synstylae*, sect. *Chinenses* and rose cultivars formed a clade within the phylogenetic tree and are further divided into several subclades. Population divergence analysis indicated that the genetic backgrounds of both sections were similar and had not yet fully differentiated. Therefore, we suggest that sect. *Synstylae* and sect. *Chinenses* can be combined and treated as different series within the same section.

Studies at the population level further confirmed the hybrid origin of certain *Rosa* accessions and clarified the phylogenetic relationships

between species. *R. foetida*, *R. platyacantha* and *R. spinosissima* have a hybrid background between sect. *Pimpinellifoliae* and sect. *Rosa*, which is consistent with previous findings[14,64,69]. According to molecular and morphological evidence, *R. pseudobanksiae* had a hybrid origin between *R. banksiae* var. *normalis* and *R. multiflora* var. *cathayensis*[70], which was corroborated in this study. The European accessions from sect. *Caninae*, sect. *Gallicanae* and *R. arvensis* formed one clade, indicating that these materials have a relatively independent genetic background compared to *Rosa* species from China. Previous studies have shown that sect. *Caninae* originated from the hybridization between *R. arvensis* and *R. majalis*[64,69]. This study also found a hybrid origin of sect. *Caninae*. However, relatively few *Rosa* samples with European backgrounds were collected in this study. Further studies should be conducted using samples from different regions across the world because the limited geographical scope of sampling may have led to differences in phylogenetic analyses[26]. These studies should combine *Rosa* species with Asian, European and North American backgrounds for a more comprehensive analysis.

In an attempt to clarify the reticulated pattern within the genus, studies have increasingly used both chloroplast genes[19,65–67] and nuclear genes[15,71,72] to construct phylogenetic trees. However, differences in topological structures caused by hybridization, introgression and linked selection are likely to be common, especially in genera that have undergone rapid divergence and polyploidization events, such as the genus *Rosa*[64,68]. In addition, contrasting biological properties of the nuclear and plastid genomes, such as their modes of inheritance, as well as recombination and evolutionary rates, may also lead to conflicting phylogenetic results[73]. Therefore, reliance on a single outcome is not sufficient for a comprehensive understanding. Based on the obtained high-quality whole genome information, we first used conserved single-copy nuclear markers, which have characteristics such as biparental inheritance and a high content of information. This tree successfully links the identified monophyletic clades with the traditional classification of the genus based on morphology. We also combined the results of plastid-based phylogeny and identified potential hybridization events. The nodes of both trees show high supporting rates (Extended Data Fig. 3). For deeper nodes, particularly within the nested subclade of sect. *Synstylae*, sect. *Chinenses* and rose cultivars, the nuclear-based phylogeny showed higher support rates compared to plastid-based analyses. In previous studies, short branch lengths with low support values were also observed in plastid phylogenetic trees[15,66], suggesting simultaneous diversification of these sections.

Tracing back the evolutionary process of *Rosa*

Combining the population history and geographical distribution of the genus *Rosa* (Extended Data Fig. 4) suggests that the northwest and southwest of China represent two independent centres of diversity of *Rosa*. Fossil and phylogenetic evidence suggest that the genus *Rosa* originated in Central Asia[15]. Based on population analysis, we believe that subg. *Hulthemia* originated in the northwest region of China, followed by differentiation and the earliest radiation of sect. *Pimpinellifoliae*[64], representing the northern populations. The species in the northern populations had yellow flowers and a larger number of leaflets with smaller leaf size, which may be the result of adaptive evolution to reduce water loss caused by high latitudes and arid climates. The southwest diversity centre of *Rosa* originated from sect. *Banksianae*, characterized by white flowers, compound corymb inflorescences and the presence of fragrance, which may be related to a warm and humid climate[74]. We speculate that this geographical pattern of flower colour may also be influenced by pollinator behaviours across different regions. In the higher-latitude areas of northwest China, the flowers of *Rosa* are mostly bright yellow or even spotted, which can attract insect pollinators more effectively. The proportion of white flowers was higher in the south of China, indicating that here, flowers may mainly rely on scent to attract insect pollinators.

The Quaternary climate change and Himalayan orogeny provided new ecological niches for the vertical distribution of plants[75,76]. Gao and colleagues analysed the population history of two species in sect. *Pimpinellifoliae*, namely, *R. sericea* and *R. omeiensis*[75,77]. Their results suggest that climate change during the Last Glacial Maximum period led to a lower tree line, thus creating additional high-altitude habitats for cold-tolerant *Rosa* species and further promoting the divergence of the two species. By combining the geographic distribution of *Rosa* species, we speculate that changes in climate and topography have promoted the differentiation of *Rosa* species along mountain ranges[77]. There is formation of an expansion trend from northern and southern evolutionary centres towards the middle area. The unique geographical environment of the Qinling region fosters the diversity of the rose populations in the central area and facilitates the genetic exchange between the northern and southern regions, thereby giving rise to the more widely distributed sect. *Rosa* and sect. *Synstylae*. Sect. *Chinenses* and sect. *Synstylae* have undergone multiple hybridizations throughout history[68]. Based on the results of this study, we believe that sect. *Chinenses* originated from sect. *Synstylae* and that the two sections have not yet fully differentiated.

Evolution of traits and identification of domestication loci

Ancestral trait reconstruction showed that the common ancestor of *Rosa* most likely showed single-petal flowers with yellow colour and seven leaflets. This inference is consistent with the number of leaflets observed in the fossil of the ancient rose *R. fortuita*[78]. The evolutionary process followed a trend of the loss of primitive yellow flowers and the emergence of diverse red/pink flowers in *Rosa*. The simultaneous selection of genes related to both carotenoid and flavonoid biosynthesis may have led to these changes. During the domestication of *Rosa* in the nineteenth and twentieth centuries, red and pink flowers were the predominant rose colours. Despite the convergence of natural and human selection in the preference for flower colour, our results indicate that these two processes may have been driven by different selection loci. For example, we identified two MYB genes that may be involved in the formation of red/pink flowers. However, they were identified in different groups and loci and may have been subject to independent selective forces. Regarding rose scent, the characteristic scent compounds differ among wild and cultivated rose materials[43,79]. This means that the gain and loss of rose scent during domestication might follow different paths compared to the evolutionary process of the genus. Both processes may lead to variations in the functions of the enzyme involved in the scent compounds in plants[79–81]. This also explains why the genes identified in this study were involved in different bio-compound syntheses, which also showed that rose scent is formed by a multitude of traits driven by many genes and pathways. However, it is worth noting that the modern rose cultivars used in this study were under-represented for rose scent. Subsequent studies should consider including more diverse germplasm to ensure a wider diversity of scent-forming compounds.

Flower colour and scent are biochemically interconnected and may be associated with traits such as flower development and responses to both biotic and abiotic stresses[46]. We expected that many enzymes and transcription factors identified in the selected loci (such as MYBs, AP2/ERFs and SPLs) might be involved in the coordinated regulation of multiple traits[10,40,82,83]. According to the results, many AP2/ERFs on different chromosomes are under selection. In apples and strawberries, AP2/ERFs have been reported to interact with MYBs during fruit de-greening and coloration[84,85]. They may also interact with TPSs to regulate volatile compounds[86,87]. In addition, the *miR156-SPL* module has been demonstrated to participate in plant secondary metabolites[88,89]. The *miR156-SPL9* identified in 'Old Blush' was reported to regulate both anthocyanin and germacrene D biosynthesis[10]. *CCoAOMT* was also involved in the regulation of anthocyanin content, leading to the differentiation of red and purple colours in rose petals[90]. These

genes may act as potential pleiotropic regulators in the evolution and domestication of roses. Nonetheless, their significance needs to be further confirmed through functional validation.

A second breeding revolution for modern roses

Over the past 500 years, ornamental plants have undergone complex and intense selective pressures[4], which may have led to the sudden loss of genetic diversity. The development of biotechnology has provided powerful tools for the genetic improvement of ornamental plants; however, the narrow genetic base still poses obstacles to breeding efforts. De novo domestication has opened up new horizons for plant breeding. On the one hand, identifying key domestication genes during the domestication process can lay the foundation for efficient genetic improvement. On the other hand, using wild relatives for re-domestication can further facilitate the introduction of broader genetic backgrounds and can overcome breeding bottlenecks, while accelerating the breeding process of desired traits[91,92].

The findings of this study suggest that the genetic diversity of cultivated roses is limited compared to genetic diversity in wild roses. This observation differs from the findings of a recent study, which reported that the genetic diversity of modern roses have a significantly higher genetic diversity than the wild roses[20]. Such discrepancies may have arisen from differential grouping of materials and the number of samples within groups. Nevertheless, given that the major contributors of modern roses are derived from only two sections, it is necessary to expand the genetic background of modern roses by introducing wild germplasm in future breeding efforts. With the aim to overcome this genetic bottleneck and initiate a secondary breeding revolution for modern roses, we have selected representative germplasm from different sections based on phylogenetic and population analyses.

First, primitive accessions with homozygous genetic backgrounds were selected (Fig. 2 and Supplementary Table 19). These accessions were differentiated early within sections, show strong resistance and can thus be used as starting materials for the re-domestication of *Rosa* plants. They can also be selected as representative germplasm for pan-genomic studies of *Rosa*.

Second, accessions that were regarded as natural intersection hybrids with heterozygous genetic backgrounds were selected as potential breeding materials (Fig. 2 and Supplementary Table 20). These materials can be further used for distant hybridization breeding to enrich the genetic composition of modern roses and expand their genetic diversity.

Methods

Sample collection

The *Rosa* accessions used in this study were extensively collected across China since 2003, resulting in a total of 205 samples (Supplementary Table 13). The collected materials were introduced and preserved in the Kunming Yang Chinese Rose Garden (Kunming, China), Xinjiang Career Technical College (Xinjiang, China) and Beijing Forestry University (Beijing, China). They were subjected to years of observation, and specimen analyses were performed. This work led to the conclusion that certain materials should not be regarded as independent species but should rather be merged with existing species (Supplementary Table 11 and Supplementary Notes). The materials covered all 12 subdivisions of *Rosa*, including three subg. *Hulthemia* samples, one subg. *Hesperhodos* sample, 65 sect. *Rosa* samples, 39 sect. *Synstylae* samples, 18 sect. *Pimpinellifoliae* samples, 17 sect. *Chinenses* samples, 12 sect. *Banksianae* samples, five sect. *Microphyllae* samples, two sect. *Laevigatae* samples, one sect. *Bracteatae* sample, four sect. *Caninae* samples, one sect. *Gallicanae* sample, 27 ancient rose samples and 20 modern rose samples. The modern rose group included 13 'Eyes for you' rose cultivars and five 'Tianshan' rose cultivars bred in China. The collected materials covered 80 species recorded in the *Flora of China*, with a coverage rate of 84%. The assignment of botanical sections of *Rosa* materials in subsequent population genetic analyses was based on the revisions proposed in Supplementary Table 11 and Supplementary Notes.

Phenotypic trait measurement

The investigation of phenotypic traits for wild *Rosa* resources has been ongoing for 20 years (since 2003). After field research, these resources were introduced to our germplasm gardens for years of observations, including the morphology of flowers, leaves and hips, as well as the shedding of glandular trichomes and sepals. All traits were measured and subsequently photographed at their full bloom stage. Measurements for all traits were taken from three biological replicates, with technical replicates performed for each quantitative trait. Binary traits included flower inflorescence (corymb or compound/solitary) and flower spot (with spot/without spot). Quantitative traits included the number of petals, RGB values of flower colour and the number of leaflets. A detailed description of the methods used to measure these traits is provided in Supplementary Methods. The detailed phenotypic information of these species is recorded in the book *Genus Rosa L. in China*[16].

DNA extraction, library preparation and sequencing

Genomic DNA was extracted from young leaves of *R. persica* using the DNeasy Plant Mini Kit (QIAGEN). DNA quality was assessed using the Agilent 4200 Bioanalyzer (Agilent Technologies) and pulsed field gel electrophoresis. The genomic DNA was subsequently processed using two library construction methods: one for long-read sequencing on the Pacific Bio platform and the other for short-read sequencing on the Illumina HiSeq platform.

A Single-molecule real-time (SMRT) bell library for long-read sequencing data was constructed using the Pacific Biosciences SMRTbell Express Template Prep kit 2.0 (Pacific Biosciences). The REV-HiFi 15 kb libraries were prepared, followed by barcode annealing (TCACGACGACGAGTAT) and binding of the SMRT bell templates to polymerases using the DNA/Polymerase Binding Kit (Pacific Biosciences). Sequencing was performed on a Pacific Biosciences Revio platform at Berry Genomics.

For short-read sequencing, the library was prepared following Illumina TruSeq Nano DNA Sample Prep Kit, and the 150 bp paired-end library was sequenced on an Illumina HiSeq X platform (Illumina). For Hi-C sequencing, fresh young leaf tissue was preserved in 1% formaldehyde (*v*/*v*). Crosslinked DNA was fixed and marked with biotin and then purified and sheared to 500–700 bp to construct sequencing libraries. Sequencing was performed using 150 bp paired-end mode on an Illumina Hiseq X platform. Petals, leaves, rose-hips and stems of the same individual were collected for RNA extraction. Total RNA was extracted using the TRIzol reagent. Libraries were generated and sequenced on an Illumina Hiseq X platform.

Complete assembly and evaluation of the *R. persica* genome

A total of 86.56 Gb (254×) PacBio HiFi reads were generated. A primary assembly was conducted using Hifiasm based on the haplotype-aware method (https://github.com/chhylp123/hifiasm). This yielded a contig-level assembly with 13 and 12 continuous sequences for haplotypes 1 and 2, respectively. The assembled contigs were subsequently anchored using Hi-C sequencing, and 55.37 Gb of raw data (163×) were obtained. Seven super-scaffolds for each haplotype, with length ranging from 36 to 65 Mb in Hap 1 and 39 to 65 Mb in Hap 2, were obtained after assigning data with ALLHIC (v.0.9.8) (https://github.com/tanghaibao/allhic). QuarTeT (v.1.2.0) (https://github.com/aaranyue/quarTeT) was used to locate and fill the remaining gaps in Hap 1 (six gaps) and Hap 2 (five gaps). The flanking sequences of these gaps were extracted, while the consistent sequences were identified by RagTag (v.2.1.0) (https://github.com/malonge/RagTag) using HiFi data and published *Rosa* genomes. Both short- and long-read sequences were used to map the genome and assess the coverage of continuous

and gap-closed sequences. It is worth noting that there was uniform coverage across the entire genome. Finally, a phased, complete genome with zero gaps was obtained.

Quality evaluation was conducted to evaluate the accuracy, continuity and completeness of the genome. Quality values were calculated using Merqury[93] (v.1.3) to evaluate the accuracy of the assembly. The completeness and continuity of the genome were evaluated using BUSCO (v.5.4.5)[94] and LAI (LTR_retriever; v.2.9.0). Samtools (v.1.9) was used to assess the genome coverage from the mapping results of Illumina and HiFi reads. Genome size evaluation and karyotype analysis, detection of telomeres and centromeres, genome annotation, quality evaluation, and comparative genomic analyses are described in Supplementary Methods.

Whole-genome re-sequencing and variant calling

A total of 205 fresh young leaves were collected for the extraction of high-quality genomic DNA. Then, a paired-end library with an insert size of 300–500 bp was constructed and detected using a Qubit 3.0 fluorometer (Life Technologies) and a Bioanalyzer 2100 system (Agilent Technologies). Sequencing of the library was performed using the Illumina NovaSeq X platform (Illumina). The clean data of 205 *Rosa* accessions used in this study and 10 previously published data were aligned to the reference genome of *R. persica* Hap 1 using BWA mem2 (v.0.7.17)[95]. Duplicated reads were then removed using the Picard (v. 2.20.7) tool MarkDuplicates (https://broadinstitute.github.io/picard/). We generated genomic variant call formats (GVCFs) for each accession using GATK (v.4.2.2.0)[96] HaplotypeCaller. The GVCFs were then combined and used for variant calling by GenotypeGVCFs. Further filtering was carried out with parameters 'MQ < 30.0, MQRankSum < −20.0, QD < 2.0, QUAL < 30.0, SOR > 3.0, FS > 60.0 and ReadPosRankSum < −8.0.'

Phylogenetic analyses

We chose four species as outgroup materials: *Fragaria nilgerrensis* (SRR10275319), *Fragaria iinumae* (SRR9217949), *Potentilla tanacetifolia* (SRR8208352) and *Rubus henryi* (SRR14240565), and the genomic data were obtained from the National Center for Biotechnology Information database. For the single-copy SNP phylogenetic tree, we used the identified Rosaceae 707 single-copy gene set and extracted genomic regions using TBtools (v.2.056)[97]. Bedtools (v.2.31.1) was then used to extract variant information from the variant call format (VCF) file based on genomic location, and a dataset of 6,048 single-copy SNPs was formed. The maximum likelihood phylogenetic tree was constructed using IQ-TREE2 (v. 2.1)[98], with a bootstrap test of 1,000 replicates. For the chloroplast phylogenetic tree, chloroplast genomes of all samples were assembled using GetOrganelle (v.1.7.4.1)[99]. PGA (v.1.9.1)[100] was used to annotate chloroplast genomes. Coding sequence (CDS) were extracted using geneious software (v.2022.2.1), and MAFFT (v.7.520) was used to perform multiple alignments on extracted sequences. The maximum likelihood phylogenetic tree was inferred using IQ-TREE2 (v. 2.1). FigTree (v.1.4.4) (http://tree.bio.ed.ac.uk/software/figtree/) and iTOL (v.7) (https://itol.embl.de) were used for visualization. The Tanglegram function in the R package 'dendextend' (v.1.18.1) was used to compare chloroplast and nuclear trees[101].

Ancestral state reconstruction

Mesquite[102] (v.3.81) was used for ancestral state reconstruction using both maximum parsimony and maximum likelihood methods. Both binary and discrete traits were used including inflorescence types (0 for solitary, 1 for corymb or compound), flower petal (0 for single petal, 1 for double petals), flower spot (0 for spotless, 1 for spotted), flower colour (0 for white, 1 for yellow, 2 for red/pink, 3 for others) and leaflet number (0 for single leaf, 1 for 3 leaflets, 2 for 5 leaflets, 3 for 7 leaflets, 4 for 9 leaflets, 5 for 11 leaflets, 6 for 13 leaflets, 7 for 15 leaflets). We also treated each value of flower colour (according to the RGB classification) as continuous traits and used the maximum likelihood method to reconstruct the ancestral RGB states of *Rosa*.

Population genetic analyses

SNPs with more than 10% missing calls and minor allele frequency (MAF) < 0.05 were removed for population genetics analyses. ADMIXTURE[103] (v. 1.3.0) was used to identify the population structure. The optimal *K* value was determined by calculating the cross-validation error. Plink[104] (v. 1.90) was used to run PCA analyses of population SNPs. PopLDdecay[105] (v. 3.41) was used for linkage disequilibrium analysis. GeneHapR (v.1.2.5)[106] was used for haplotype analysis.

GWAS

GWAS analysis was performed on the number of leaflets (quantitative), the number of petals (quantitative) and inflorescence type (binary). Only diploid and tetraploid samples were included in the GWAS analysis using GWASpoly software (v.2.13)[107]. A total of 861,222 SNPs were used for initial filtering. SNPs with more than 10% missing rate and MAF < 0.05 were removed, which resulted in a set of 96,342 markers for subsequent analysis. The VCF2dosage function was first used to separately generate dosage files for both diploids and tetraploids from VCF files. The marker dosage was then combined and coded based on the highest ploidy. The (0, 1, 2) dosage for diploids were coded as (0, 2, 4). The ploidy flag was set as 'ploidy = 4'. The leave-one-chromosome-out method was used to correct the population structure. The Bonferroni test was used to determine the logarithm of the odds threshold, corresponding to a 5% false-positive rate across the genome. The get.QTL function was used to provide an output of the obtained significant markers, with parameters 'bp.window = 5e6' set to prune the marker list. Manhattan plots were generated using the CMPlot package (v.4.5.1) (https://github.com/YinLiLin/CMplot) in R (v.4.3.1). Candidate genes were screened within the 50 kb regions upstream and downstream of the peak SNP.

Selective sweep analysis

Data of polyploid samples were first removed using VCFtools (v.0.1.16)[108], leaving 149 diploid samples for subsequent analysis. Selective sweeps were then identified using whole-genome SNPs between early-differentiated (group 1) and recently differentiated (group 2) samples, as well as between wild materials (group 2-W) and cultivars (group 2-C) within group 2. A sliding window size of 100 kb with a step size of 10 kb was used to calculate F_{ST} and nucleotide diversity (π) between groups. The top 5% regions of both F_{ST} values and π ratios were considered as candidate selective sweeps and were used to identify candidate genes.

Geographical distribution and gene flow analysis

The distribution information of species in the genus *Rosa* in China was derived from field investigation records, specimen collection information and geographic distribution records in the *Flora of China*. The collected information was then converted into Global Positioning System coordinate data, and the distribution map (Extended Data Fig. 4) was drawn using ArcGIS (v.10.2) software.

For gene flow analyses, polyploid species were excluded. TreeMix[109] (v.1.13) was used to identify gene flows between different subdivisions of the genus *Rosa*. To obtain the most suitable migration edges, the number of migrations was set from 0 to 10. The OptM R package (v.0.1.6; https://cran.r-project.org/web/packages/OptM) was used to identify the best migration edges. *D*-statistics and f4 statistics were calculated using the Dtrios command in Dsuite (v.0.4 r42)[110] software to identify genome-wide introgressions based on high-quality SNPs.

Demographic analysis

To examine the recent demographic history of *Rosa*, we used the easySFS script (available at https://github.com/isaacovercast/easySFS) to convert VCF files into an SNP frequency spectrum format. This procedure was followed by the application of Stairway Plot[111] software (v.2.1),

wherein a random subset comprising 67% of all loci was extracted. This step facilitates the inference of temporal shifts in the effective population size (N_e) based on neutral loci. The mutation rate of *Rosa* was estimated based on fourfold degenerate sites that were neutral to selection pressure. Seven different nucleotide substitution models (models 1–7) were used in PAML (v.X1.3.1)[112]. An estimated neutral mutation rate of 2.9×10^{-9} was obtained by calculating the average results. It is worth noting that this mutation rate is also supported by published references[35,75]. The generational time span for plants in the genus *Rosa* was established at 2 years.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Raw sequencing data have been deposited in the National Center for Biotechnology Information (NCBI) BioProject database with the BioProject PRJNA1224503. The phased genome assembly of *R. persica* has been deposited in the NCBI database under accession numbers JBLIYO000000000 (Hap 1) and JBLIYP000000000 (Hap 2). The previously published reads used in this study are available from NCBI and public GDR database at https://www.rosaceae.org. Specific source data are listed in Supplementary Tables 10 and 13.

References

1. Altman, A., Shennan, S. & Odling-Smee, J. Ornamental plant domestication by aesthetics-driven human cultural niche construction. *Trends Plant Sci.* **27**, 124–138 (2022).
2. Dempewolf, H. et al. Past and future use of wild relatives in crop breeding. *Crop Sci.* **57**, 1070–1082 (2017).
3. Cheng, S. et al. Harnessing landrace diversity empowers wheat breeding. *Nature* **632**, 823–831 (2024).
4. Bohra, A. et al. Reap the crop wild relatives for breeding future crops. *Trends Biotechnol.* **40**, 412–431 (2022).
5. Smulders, M. J. M., et al. In *Wild Crop Relatives: Genomic and Breeding Resources: Plantation and Ornamental Crops* (ed. Kole, C.) 243–275 (Springer, 2011).
6. Krüssmann, G. *The Complete Book of Roses* (Timber, 1981).
7. Thibault, L., et al. Dark side of the honeymoon: reconstructing the Asian x European rose breeding history through the lens of genomics. Preprint at *bioRxiv*, https://doi.org/10.1101/2023.06.22.546162 (2024).
8. Purugganan, M. D. What is domestication? *Trends Ecol. Evol.* **37**, 663–671 (2022).
9. Wyile, A. The history of garden roses. Part I. *J. Roy. Hort. Soc.* **79**, 555–571 (1954).
10. Raymond, O. et al. The *Rosa* genome provides new insights into the domestication of modern roses. *Nat. Genet.* **50**, 772–777 (2018).
11. Tan, J. et al. Genetic relationships and evolution of old Chinese garden roses based on SSRs and chromosome diversity. *Sci. Rep.* **7**, 15437 (2017).
12. Vukosavljev, M. et al. Genetic diversity and differentiation in roses: a garden rose perspective. *Sci. Hortic.* **162**, 320–332 (2013).
13. Matsumoto, S. et al. Phylogenetic analyses of the genus *Rosa* using the matK sequence: molecular evidence for the narrow genetic background of modern roses. *Sci. Hortic.* **77**, 73–82 (1998).
14. Koopman, W. J. M. et al. AFLP markers as a tool to reconstruct complex relationships: a case study in *Rosa* (Rosaceae). *Am. J. Bot.* **95**, 353–366 (2008).
15. Fougere-Danezan, M., Joly, S., Bruneau, A., Gao, X. F. & Zhang, L. B. Phylogeny and biogeography of wild roses with specific attention to polyploids. *Ann. Bot.* **115**, 275–291 (2015).
16. Luo, L. *Genus Rosa L. in China* (China Forestry Publishing House, 2024).
17. Rehder, A. *Bibliography of Cultivated Trees and Shrubs Hardy in the Cooler Temperate Regions of the Northern Hemisphere* (Arnold Arboretum of Harvard University, 1949).
18. Wissemann, V. *Reference Module in Life Sciences* (Elsevier, 2017).
19. Wissemann, V. & Ritz, C. M. The genus *Rosa* (Rosoideae, Rosaceae) revisited: molecular analysis of nrITS-1 and atp B-rbc L intergenic spacer (IGS) versus conventional taxonomy. *Bot. J. Linn. Soc.* **147**, 275–290 (2005).
20. Zhang, Z. et al. Haplotype-resolved genome assembly and resequencing provide insights into the origin and breeding of modern rose. *Nat. Plants* **10**, 1659–1671 (2024).
21. Zhang, X. et al. Haplotype-resolved genome assembly of the diploid *Rosa chinensis* provides insight into the mechanisms underlying key ornamental traits. *Mol. Hortic.* **4**, 14 (2024).
22. Zhong, M. et al. Rose without prickle: genomic insights linked to moisture adaptation. *Natl Sci. Rev.* **8**, nwab092 (2021).
23. Chen, F. et al. A chromosome-level genome assembly of rugged rose (*Rosa rugosa*) provides insights into its evolution, ecology, and floral characteristics. *Hortic. Res.* **8**, 141 (2021).
24. Hibrand Saint-Oyant, L. et al. A high-quality genome sequence of *Rosa chinensis* to elucidate ornamental traits. *Nat. Plants* **4**, 473–484 (2018).
25. Ku, T. R. K. *Rosa* (Rosaceae). In *Flora of China 9* (eds Wu, Z.Y. & Raven, P.H.) (Missouri Botanical Garden, 2003).
26. Schanzer, I. A., Fedorova, A. V. & Meschersky, I. G. A haplotype network approach to reconstruct the phylogeny of *Rosa* L. (Rosaceae). *Biol. Bull.* **51**, 331–345 (2024).
27. Lin, Y. Z. et al. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic. Res.* **10**, uhad127 (2023).
28. Springer, N. M. Transposable elements: microbiomes in the genomes. *Nat. Plants* **1**, 15004 (2015).
29. Bennetzen, J. L. & Wang, H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu. Rev. Plant Biol.* **65**, 505–530 (2014).
30. Tang, Y.-W. et al. *Rosa forrestiana* var. *maculata*, a new variety of *Rosa* (Rosaceae) from Yunnan, China. *Phytotaxa* **652**, 4 (2024).
31. Zheng, L.-N. et al. *Rosa funingensis* (Rosaceae), a new species from Yunnan, China. *PhytoKeys* **229**, 61–70 (2023).
32. Lyu, P. et al. *Rosa yangii* (Rosaceae), a new species from China. *Kew Bull.* **78**, 663–671 (2023).
33. Deng, T. et al. *Rosa tomurensis*, a new species of *Rosa* (Rosaceae) from China. *Phytotaxa* **556**, 169–177 (2022).
34. Liorzou, M. et al. Nineteenth century French rose *Rosa* sp. germplasm shows a shift over time from a European to an Asian genetic background. *J. Exp. Bot.* **67**, 4711–4725 (2016).
35. Zang, F. et al. Resequencing of *Rosa rugosa* accessions revealed the history of population dynamics, breed origin, and domestication pathways. *BMC Plant Biol.* **23**, 235 (2023).
36. Ao, H. et al. Global warming-induced Asian hydrological climate transition across the Miocene–Pliocene boundary. *Nat. Commun.* **12**, 6935 (2021).
37. Clark, P. U. et al. The Last Glacial Maximum. *Science* **325**, 710–714 (2009).
38. Ogata, J., Kanno, Y., Itoh, Y., Tsugawa, H. & Suzuki, M. Anthocyanin biosynthesis in roses. *Nature* **435**, 757–758 (2005).
39. Fukuchi-Mizutani, M. et al. Biochemical and molecular characterization of anthocyanidin/flavonol 3-glucosylation pathways in *Rosa× hybrida*. *Plant Biotechnol.* **28**, 239–244 (2011).
40. He, G., Zhang, R., Jiang, S., Wang, H. & Ming, F. The MYB transcription factor RcMYB1 plays a central role in rose anthocyanin biosynthesis. *Hortic. Res.* **10**, uhad080 (2023).

41. Li, M. et al. Rosa1, a transposable element-like insertion, produces red petal coloration in rose through altering RcMYB114 transcription. *Front. Plant Sci.* **13**, 857684 (2022).
42. Shen, Y. et al. RrMYB5- and RrMYB10-regulated flavonoid biosynthesis plays a pivotal role in feedback loop responding to wounding and oxidation in *Rosa rugosa*. *Plant Biotechnol. J.* **17**, 2078–2095 (2019).
43. Guterman, I. et al. Rose scent: genomics approach to discovering novel floral fragrance-related genes. *Plant Cell* **14**, 2325–2338 (2002).
44. Spiller, M., Berger, R. G. & Debener, T. Genetic dissection of scent metabolic profiles in diploid rose populations. *Theor. Appl. Genet.* **120**, 1461–1471 (2010).
45. Sakai, M. et al. Production of 2-phenylethanol in roses as the dominant floral scent compound from L-phenylalanine by two key enzymes, a PLP-dependent decarboxylase and a phenylacetaldehyde reductase. *Biosci. Biotechnol. Biochem.* **71**, 2408–2419 (2007).
46. Li, Y. et al. The coordinated interaction or regulation between floral pigments and volatile organic compounds. *Hortic. Plant J.* https://doi.org/10.1016/j.hpj.2024.01.002 (2024).
47. Li, H. et al. The complexity of volatile terpene biosynthesis in roses: particular insights into β-citronellol production. *Plant Physiol.* **196**, 1908–1922 (2024).
48. Soufflet-Freslon, V. et al. Diversity and selection of the continuous-flowering gene, *RoKSN*, in rose. *Hortic. Res.* **8**, 76 (2021).
49. Iwata, H. et al. The TFL1 homologue KSN is a regulator of continuous flowering in rose and strawberry. *Plant J.* **69**, 116–125 (2012).
50. Dugo, M. L. et al. Genetic mapping of QTLs controlling horticultural traits in diploid roses. *Theor. Appl. Genet.* **111**, 511–520 (2005).
51. Shubin, L. et al. Inheritance of perpetual blooming in *Rosa chinensis* 'Old Blush'. *Hortic. Plant J.* **1**, 108–112 (2015).
52. Kania, T., Russenberger, D., Peng, S., Apel, K. & Melzer, S. *FPF1* promotes flowering in *Arabidopsis*. *Plant Cell* **9**, 1327–1338 (1997).
53. Lei, Y. et al. Woodland strawberry WRKY71 acts as a promoter of flowering via a transcriptional regulatory cascade. *Hortic. Res.* **7**, 137 (2020).
54. Kim, H.-J. et al. A genetic link between cold responses and flowering time through FVE in *Arabidopsis thaliana*. *Nat. Genet.* **36**, 167–171 (2004).
55. Roman, H. et al. Genetic analysis of the flowering date and number of petals in rose. *Tree Genet. Genomes* **11**, 85 (2015).
56. François, L. et al. A miR172 target-deficient AP2-like gene correlates with the double flower phenotype in roses. *Sci. Rep.* **8**, 12912 (2018).
57. Gattolin, S. et al. Deletion of the miR172 target site in a TOE-type gene is a strong candidate variant for dominant double-flower trait in Rosaceae. *Plant J.* **96**, 358–371 (2018).
58. Lopez Arias, D. C. et al. Characterization of black spot resistance in diploid roses with QTL detection, meta-analysis and candidate-gene identification. *Theor. Appl. Genet.* **133**, 3299–3321 (2020).
59. Li, X. R. et al. Function of two splicing variants of RcCPR5 in the resistance of *Rosa chinensis* to powdery mildew. *Plant Sci.* **335**, 111678 (2023).
60. Rawandoozi, Z. J. et al. QTL mapping and characterization of black spot disease resistance using two multi-parental diploid rose populations. *Hortic. Res.* **10**, uhad059 (2023).
61. Radauer, C., Lackner, P. & Breiteneder, H. The Bet v 1 fold: an ancient, versatile scaffold for binding of large, hydrophobic ligands. *BMC Evol. Biol.* **8**, 286 (2008).
62. Li, R. et al. Integrated proteomic analysis reveals interactions between phosphorylation and ubiquitination in rose response to *Botrytis* infection. *Hortic. Res.* **11**, uhad238 (2024).
63. Linnaeus, C. *Species Plantarum, Holmiae: Impensis Laurentii Salvii*, Vol. 1, 491–492 (1753).
64. 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 (2022).
65. Cui, W. H. et al. Complex and reticulate origin of edible roses (*Rosa*, Rosaceae) in China. *Hortic. Res.* **9**, uhab051 (2022).
66. Zhang, C., Li, S.-Q., Xie, H.-H., Liu, J.-Q. & Gao, X.-F. Comparative plastid genome analyses of *Rosa*: insights into the phylogeny and gene divergence. *Tree Genet. Genomes* **18**, 20 (2022).
67. Bruneau, A., Starr, J. R. & Joly, S. Phylogenetic relationships in the genus *Rosa*: new evidence from chloroplast DNA sequences and an appraisal of current knowledge. *Syst. Bot.* **32**, 366–378 (2007).
68. Zhu, Z.-M., Gao, X.-F. & Fougère-Danezan, M. Phylogeny of *Rosa* sections *Chinenses* and *Synstylae* (Rosaceae) based on chloroplast and nuclear markers. *Mol. Phylogenet. Evol.* **87**, 50–64 (2015).
69. Liu, C. Y. et al. Phylogenetic relationships in the genus *Rosa* revisited based on rpl16, trnL-F, and atpB-rbcL sequences. *Hortscience* **50**, 1618–1624 (2015).
70. Zhang, C. et al. Molecular and morphological evidence for hybrid origin and matroclinal inheritance of an endangered wild rose, *Rosa*×*pseudobanksiae* (Rosaceae) from China. *Conserv. Genet.* **21**, 1–11 (2020).
71. Meng, J., Fougère-Danezan, M., Zhang, L.-B., Li, D.-Z. & Yi, T.-S. Untangling the hybrid origin of the Chinese tea roses: evidence from DNA sequences of single-copy nuclear and chloroplast genes. *Plant Syst. Evol.* **297**, 157–170 (2011).
72. Debray, K. et al. Identification and assessment of variable single-copy orthologous (SCO) nuclear loci for low-level phylogenomics: a case study in the genus *Rosa* (Rosaceae). *BMC Evol. Biol.* **19**, 152 (2019).
73. Zuntini, A. R. et al. Phylogenomics and the rise of the angiosperms. *Nature* **629**, 843–850 (2024).
74. Wang, Q., Guo, Q., Chi, X., Zhu, S. & Tang, Z. Evolutionary history and climate conditions constrain the flower colours of woody plants in China. *J. Plant Ecol.* **15**, 196–207 (2021).
75. Gao, Y.-D., Zhang, Y., Gao, X.-F. & Zhu, Z.-M. Pleistocene glaciations, demographic expansion and subsequent isolation promoted morphological heterogeneity: a phylogeographic study of the alpine *Rosa sericea* complex (Rosaceae). *Sci. Rep.* **5**, 11698 (2015).
76. Xing, Y. & Ree, R. H. Uplift-driven diversification in the Hengduan Mountains, a temperate biodiversity hotspot. *Proc. Natl Acad. Sci. USA* **114**, E3444–E3451 (2017).
77. Gao, Y.-D., Gao, X.-F. & Harris, A. Species boundaries and parapatric speciation in the complex of alpine shrubs, *Rosa sericea* (Rosaceae), based on population genetics and ecological tolerances. *Front. Plant Sci.* **10**, 321 (2019).
78. Su, T. et al. A Miocene leaf fossil record of *Rosa* (*R. fortuita* n. sp.) from its modern diversity center in SW China. *Palaeoworld* **25**, 104–115 (2016).
79. Caissard, J. C., Adrar, I., Conart, C., Paramita, S. N. & Baudino, S. Do we really know the scent of roses? *Bot. Lett.* https://doi.org/10.1080/23818107.2022.2160807 (2022).
80. Magnard, J. L. et al. Biosynthesis of monoterpene scent compounds in roses. *Science* **349**, 81–83 (2015).
81. Sun, P., Schuurink, R. C., Caissard, J. C., Hugueney, P. & Baudino, S. My way: noncanonical biosynthesis pathways for plant volatiles. *Trends Plant Sci.* **21**, 884–894 (2016).

82. Zvi, M. M. et al. PAP1 transcription factor enhances production of phenylpropanoid and terpenoid scent compounds in rose flowers. *New Phytol.* **195**, 335–345 (2012).
83. Liu, J. Y., Osbourn, A. & Ma, P. D. MYB transcription factors as regulators of phenylpropanoid metabolism in plants. *Mol. Plant* **8**, 689–708 (2015).
84. Zhai, Y. et al. APETALA2/ethylene responsive factor in fruit ripening: roles, interactions and expression regulation. *Front. Plant Sci.* **13**, 979348 (2022).
85. Dang, Q. Y. et al. An apple (*Malus domestica*) AP2/ERF transcription factor modulates carotenoid accumulation. *Hortic. Res.* **8**, 223 (2021).
86. Shen, S. L. et al. CitAP2.10 activation of the terpene synthase CsTPS1 is associated with the synthesis of (+)-valencene in 'Newhall' orange. *J. Exp. Bot.* **67**, 4105–4115 (2016).
87. Wei, C. Y. et al. Linalool synthesis related PpTPS1 and PpTPS3 are activated by transcription factor PpERF61 whose expression is associated with DNA methylation during peach fruit ripening. *Plant Sci.* **317**, 111200 (2022).
88. Cui, L. G., Shan, J. X., Shi, M., Gao, J. P. & Lin, H. X. The miR156-SPL9-DFR pathway coordinates the relationship between development and abiotic stress tolerance in plants. *Plant J.* **80**, 1108–1117 (2014).
89. Yu, Z.-X. et al. Progressive regulation of sesquiterpene biosynthesis in *Arabidopsis* and patchouli (*Pogostemon cablin*) by the miR156-targeted SPL transcription factors. *Mol. Plant* **8**, 98–110 (2015).
90. Zhu, Z. Q. et al. An O-methyltransferase gene, *RrCCoAOMT1*, participates in the red flower color formation of *Rosa rugosa*. *Sci. Hortic.* **336**, 113402 (2024).
91. Yu, H. & Li, J. Y. Breeding future crops to feed the world through de novo domestication. *Nat. Commun.* **13**, 1171 (2022).
92. Zsögön, A. et al. De novo domestication of wild tomato using genome editing. *Nat. Biotechnol.* **36**, 1211–1216 (2018).
93. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
94. 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**, 3210–3212 (2015).
95. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
96. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
97. Chen, C. J. et al. TBtools: an integrative toolkit developed for interactive analyses of big biological data.*Mol. Plant* **13**, 1194–1202 (2020).
98. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
99. Jin, J. J. et al. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* **21**, 241 (2020).
100. Qu, X. J., Moore, M. J., Li, D. Z. & Yi, T. S. PGA: a software package for rapid, accurate, and flexible batch annotation of plastomes. *Plant Methods* **15**, 50 (2019).
101. Galili, T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* **31**, 3718–3720 (2015).
102. Maddison, W. P. and Maddison, D. R. Mesquite: a modular system for evolutionary analysis. Version 3.81. (2023); http://www.mesquiteproject.org .
103. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
104. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
105. Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M. & Yang, T.-L. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35**, 1786–1788 (2018).
106. Zhang, R. L., Jia, G. Q. & Diao, X. M. geneHapR: an R package for gene haplotypic statistics and visualization. *BMC Bioinformatics* **24**, 199 (2023).
107. Rosyara, U. R., De Jong, W. S., Douches, D. S. & Endelman, J. B. Software for genome-wide association studies in autopolyploids and its application to potato. *Plant Genome* **9**, plantgenome2015.08.0073 (2016).
108. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
109. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).
110. Malinsky, M., Matschiner, M. & Svardal, H. Dsuite - fast D-statistics and related admixture evidence from VCF files. *Mol. Ecol. Resour.* **21**, 584–595 (2021).
111. Liu, X. & Fu, Y. X. Stairway Plot 2: demographic history inference with folded SNP frequency spectra. *Genome Biol.* **21**, 280 (2020).
112. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).

Acknowledgements

We thank Y. Yang (Kunming Yang Chinese Rose Gardening) for collecting wild *Rosa* resources and for helpful discussions on *Rosa* taxonomy. We are grateful to Y. Sui and R. Guo (Xinjiang Career Technical College) for their efforts in collecting *Rosa* resources in Xinjiang. We thank X. Gao (Key Laboratory of Molecular Epigenetics of Ministry of Education (MOE), Northeast Normal University) for his inspiring discussion and comments. We appreciate the valuable suggestions on the manuscript provided by R. Smulders and P. Arens (Plant Breeding, Wageningen University and Research). We thank J. Endelman (University of Wisconsin) for his instruction on GWASpoly software. We thank Man Zhang and T. Zheng for the suggestions on improving the paper. We thank J. Zhao (Boyce Thompson Institute) for his instruction on demographic analyses. We thank Yuxuan Zhang, L. Ji, Yujing Zhang, L. Jiang, C. Feng, Y. Zhuang, Z. Ou, R. Wang, J. Tang and K. Xiong for their assistance with the investigations of *Rosa* species. Furthermore, we thank Wuhan Dazhong Yuansheng Technology, Wuhan Feisha Genetic Information (www.frasergen.com) and BerryGenomics (www.berrygenomics.com) for providing essential sequencing service. Finally, the authors are profoundly indebted to every member of the Fishpond family, whose collective efforts were instrumental in the fruition of this work. This research was supported by the Fundamental Research Funds for the Central Universities (grant number QNTD202306 to C. Yu), National Key R&D Program of China (grant number 2019YFD1000400 to C. Yu) and National Natural Science Foundation of China (grant numbers 32471955 and 32071818 to C. Yu).

Author contributions

C. Yu designed and managed the project. B.C. performed population genomic analyses and wrote the manuscript. K.Z. assembled the *Rosa persica* genome and performed genomic analyses. M.Z. performed GWAS analysis. P.M.B. provided guidance on GWAS analysis and revised the manuscript. L. Zhou and Y.S. managed the plant materials. S.W. performed population history analyses. L.G. performed selective sweep analyses. W.D. performed demographic analyses. C. Yang and

J.C. constructed phylogenetic trees and performed phylogenetic analyses. R.H. designed and revised the figures. X.T. performed karyotype analyses of *Rosa* materials. L. Zhang performed data analyses. C. Yu, B.C., K.Z., M.Z., L.G., W.D. and S.W. wrote the paper with contributions from H.H., Y.H., L.L., H.P. and Q.Z.

Competing interests
The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41477-025-01955-5.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41477-025-01955-5.

Correspondence and requests for materials should be addressed to Chao Yu.

Peer review information *Nature Plants* thanks Xingtan Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

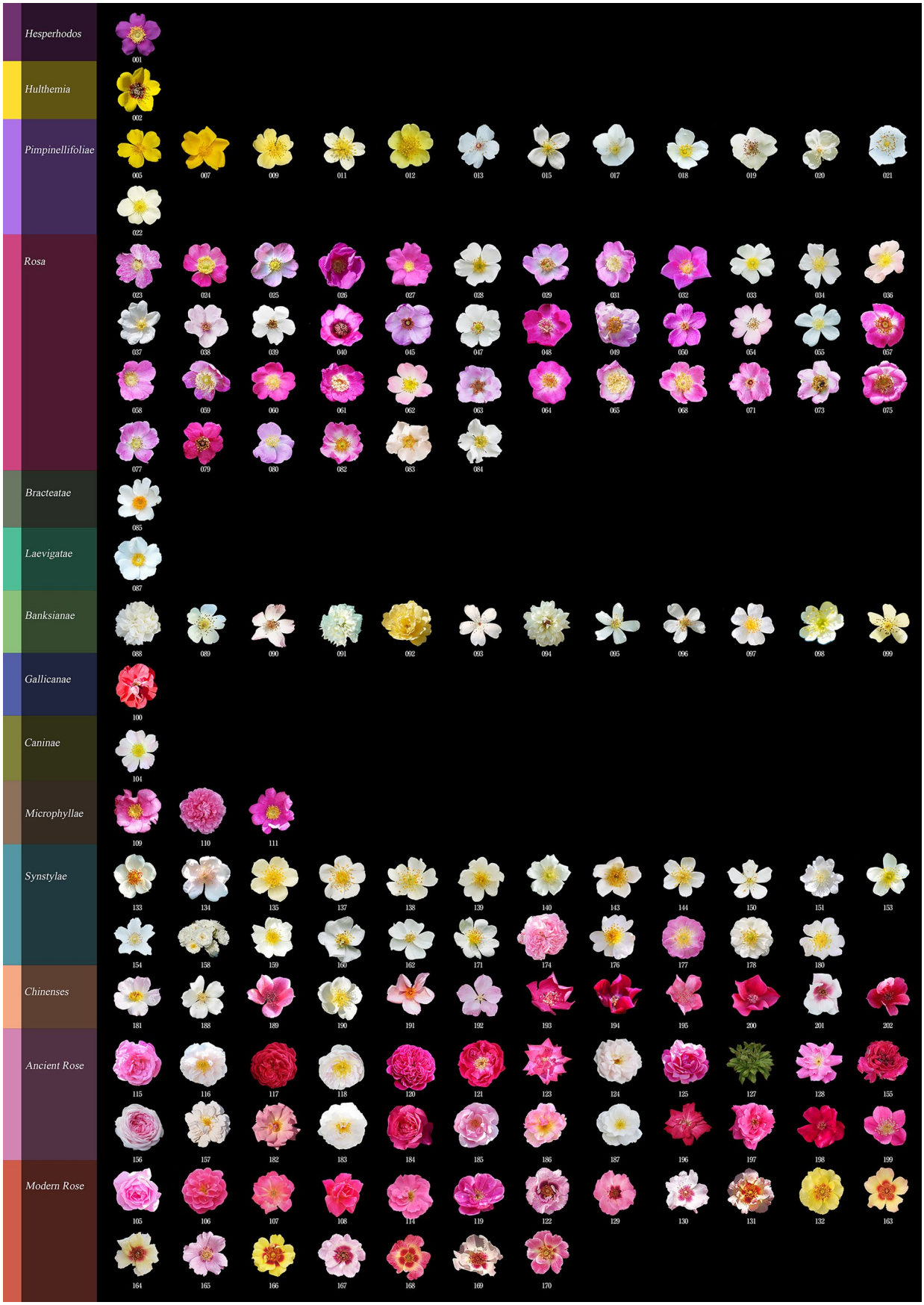
Reprints and permissions information is available at www.nature.com/reprints.

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

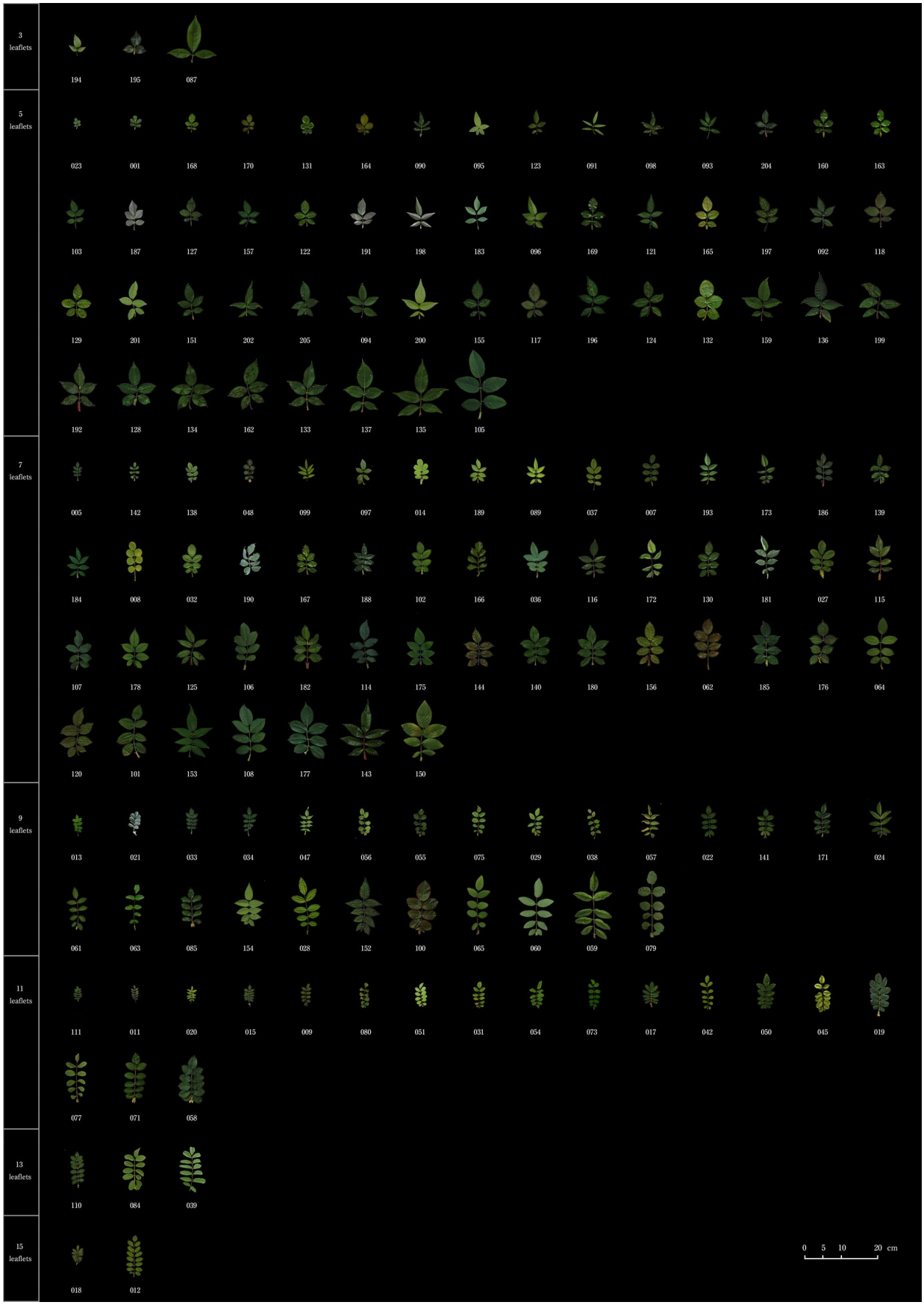
© The Author(s), under exclusive licence to Springer Nature Limited 2025

https://doi.org/10.1038/s41477-025-01955-5



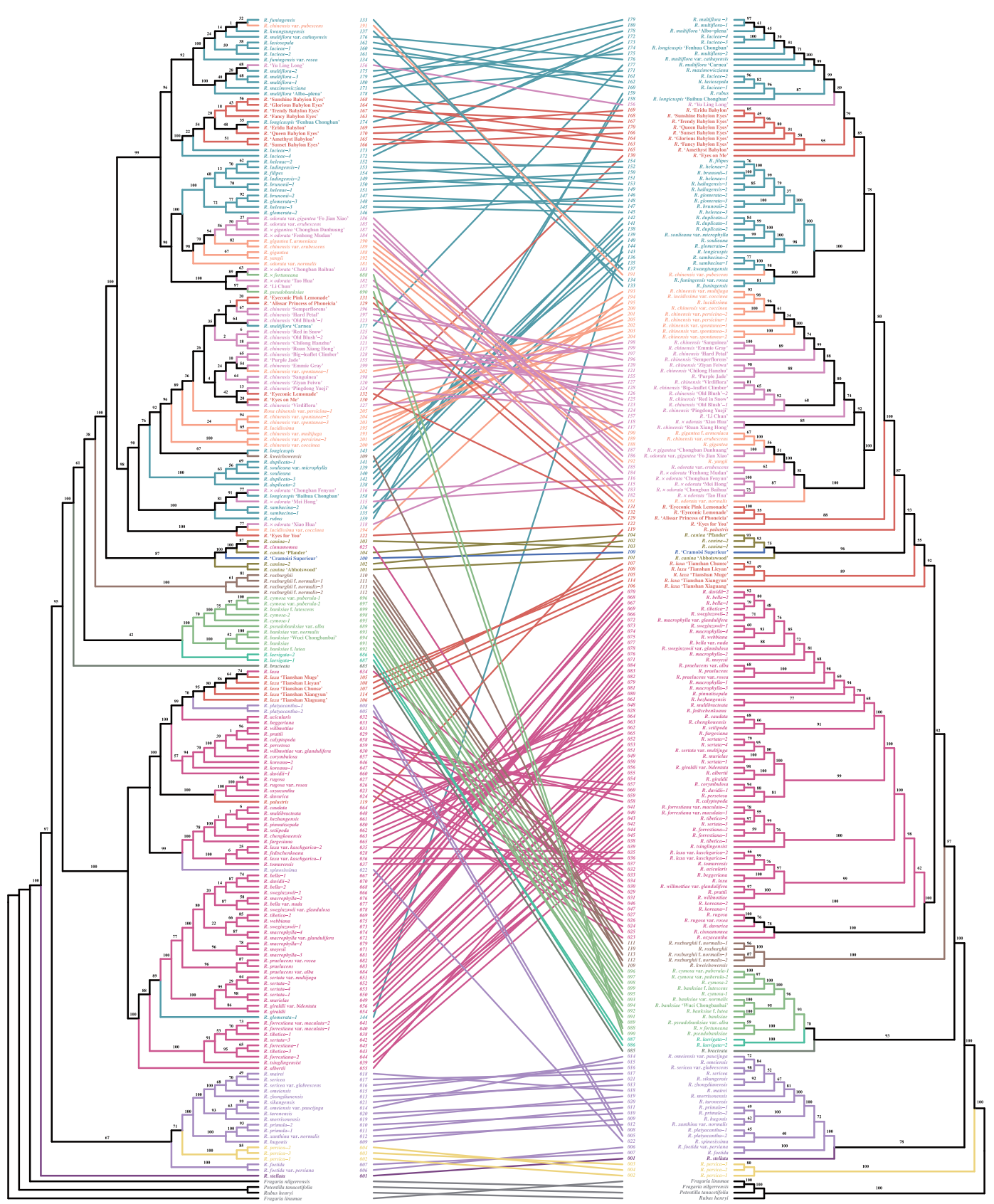
Extended Data Fig. 1 | Flower traits of *Rosa* accessions used in this study. The plates of accessions are sorted by scientific groups. Sample IDs correspond to Supplementary Table 13.

https://doi.org/10.1038/s41477-025-01955-5



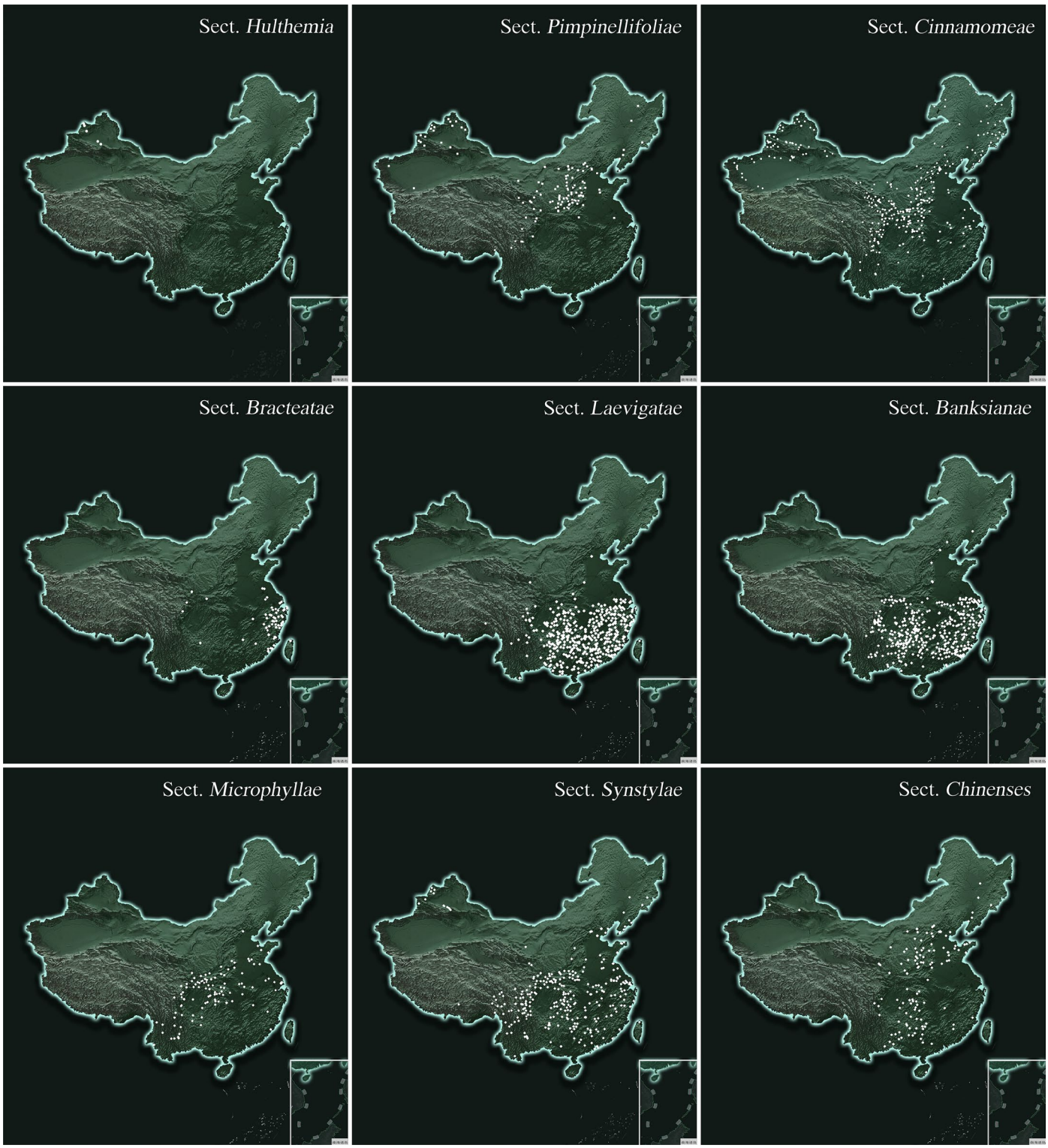
Extended Data Fig. 2 | Leaf traits of *Rosa* accessions used in this study. The plates of accessions are sorted by number of leaflets. Sample IDs correspond to Supplementary Table 13.

https://doi.org/10.1038/s41477-025-01955-5



Extended Data Fig. 3 | Comparison between plastid phylogeny (left) based on chloroplast coding sequences, and nuclear phylogeny (right) based on single-copy SNPs. Both trees were constructed using maximum likelihood method. Bootstrap values were tested with 1,000 replicates. The colors represent accessions from different botanical groups, in correspondence with previous figures.

https://doi.org/10.1038/s41477-025-01955-5



Extended Data Fig. 4 | Geographical distributions of different botanical sections of *Rosa*. The white dots represent distribution information of *Rosa* accessions, which was derived from field investigation, specimen information, and geographic distribution records from Flora of China and Chinese Virtual Herbarium (http://www.cvh.ac.cn).

nature portfolio

Corresponding author(s): Chao Yu

Last updated by author(s): Feb 7, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software had been used for data collection. Data were sequenced from PacBio and Illumina platform.
Data analysis	Hifiasm (v0.16.1), HiCUP (v0.6.1), ALLHIC (v0.9.8), Juicebox (v.1.11.08), Merqury (v.1.3), BWA mem2 (v.0.7.17), Picard (v. 2.20.7), GATK (v. 4.2.2.0), TBtools (v2.056), IQ-TREE2 (v. 2.1), GetOrganelle (v. 1.7.4.1), FigTree, iTOL, Mesquite (v.3.81), ADMIXTURE (v. 1.3.0), Plink (v. 1.90), PopLDdecay (v. 3.41), GeneHapR, GWASpoly (v. 2.10), VCFtools (v.0.1.16), ArcGIS (v.10.2), TreeMix (v.1.13), OptM R (v.0.1.6), Dsuite, Stairway Plot (v.2.1), R (v4.3.1), JCVI, QuarTeT, Genewise (v. 2.4.1), Augustus (v. 3.3), TopHat (v. 2.1.1), EVidenceModeler (v. 1.1.1), RepeatMasker (v.4.0.6), RepeatModeler (v.1.0.11), LTR-FINDER (v1.0.7), NGenomeSyn, Orthofinder (v 2.5.4), BUSCO (v.5.4.5). Specific parameters used during run-time are provided in the methods. All softwares or scripts are available from official websites or GitHub as indicated in the methods and supplementary methods.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The phased genome assembly of R. persica has been deposited in the National Center for Biotechnology Information BioProject database with the BioProject PRJNA1216766 (Haplotype 1) and PRJNA1216767 (Haplotype 2). The previously published reads used in this study are available from NCBI and public GDR database (https://www.rosaceae.org). Specific source data are listed in the Supplementary Tables 10 and 13.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Not applicable.
Reporting on race, ethnicity, or other socially relevant groupings	Not applicable.
Population characteristics	Not applicable.
Recruitment	Not applicable.
Ethics oversight	Not applicable.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	215 accessions were used to represent species of the genus Rosa. The logic of this selection was based on Rosa species that are collectible. Our collection encompasses 84% (80/95) of Rosa species documented in the Flora of China. We have also included subspecies, varieties, and cultivars for comprehensive analyses. No statistical methods were used to predetermine sample size. Our samples were all from wild type and did not use processed samples and groups.
Data exclusions	No data were excluded.
Replication	The genome sequence was taken and sequenced with more than 247 fold coverage. Robustness of phylogenetic analysis was tested using bootstrap test. All measurements of the phenotypic traits were taken at least three biological repetition. Multiple tests were used to identify gene flow, including TreeMix and D statistics. We confirm that all attempts at replication were successful.
Randomization	No random sampling is required for genome sequencing, because the genome differences are very small within the wild population, thus any wild plant is allowed for genome sequencing.
Blinding	Blinding is not applicable in our study because it does not involve subjects which receive different treatments.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☐	☒ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes
☒	☐ Public health
☒	☐ National security
☒	☐ Crops and/or livestock
☒	☐ Ecosystems
☒	☐ Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes
☒	☐ Demonstrate how to render a vaccine ineffective
☒	☐ Confer resistance to therapeutically useful antibiotics or antiviral agents
☒	☐ Enhance the virulence of a pathogen or render a nonpathogen virulent
☒	☐ Increase transmissibility of a pathogen
☒	☐ Alter the host range of a pathogen
☒	☐ Enable evasion of diagnostic/detection modalities
☒	☐ Enable the weaponization of a biological agent or toxin
☒	☐ Any other potentially harmful combination of experiments and agents

Plants

Seed stocks	The plant accessions used in this study were extensively collected across China and preserved in germplasm gardens in Yunnan, China.
Novel plant genotypes	No novel plant genotypes were applied in this study.
Authentication	We introduced these resources to our germplasm gardens for years of observations, including the morphology of flowers, leaves and hips, as well as the shedding of glandular trichomes and sepals. After comparing with a large number of specimens and consulting with experts and colleagues, the collected plant accessions were accurately identified and used for this study.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Fresh leaves of Rosa persica were collected and chopped, then put into dissociation solution for 10 min, followed by filtering to collect nucleus cells. Leaf material of Solanum pimpinellifolium was used as an internal standard, with known genome size, and co-chopped with the rose callus.
Instrument	BD FACScalibur flow cytometry (Becton,Dickinson and Company)
Software	Modifit 3.0 was used for DNA content analysis.
Cell population abundance	The fluorescence intensity was detected under 488 nm blue light excitation. For every analysis 10,000 nuclei were analyzed. Genome sizes were calculated based on the ratio of the peak position of the Rosa persica callus material and the Solanum pimpinellifolium with the known genome size (900Mbp).
Gating strategy	No gating was applied. The analysis of the plant genome size results in a non-gated one parameter histogram output on a linear scale.

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.