

Haplotype-resolved genome assembly and resequencing provide insights into the origin and breeding of modern rose

Received: 21 December 2023

Accepted: 13 September 2024

Published online: 11 October 2024

 Check for updates

Zhao Zhang  ^{1,17}, Tuo Yang  ^{1,17}, Yang Liu  ^{1,17}, Shan Wu  ^{2,17}, Honghe Sun ^{2,3,17}, Jie Wu ^{1,17}, Yonghong Li ^{4,17}, Yi Zheng  ^{5,6}, Haoran Ren ¹, Yuyong Yang ⁷, Shaochuan Shi ¹, Wenyan Wang ¹, Qi Pan ¹, Lijuan Lian ⁸, Shaowen Duan ⁹, Yingxiong Zhu ¹⁰, Youming Cai ¹¹, Hougao Zhou ¹², Hao Zhang ¹³, Kaixue Tang ¹³, Jiaopeng Cui ¹⁴, Dan Gao  ¹⁵, Liyang Chen  ¹⁵, Yunhe Jiang  ¹, Xiaoming Sun  ¹, Xiaofeng Zhou  ¹, Zhangjun Fei  ^{2,16}✉, Nan Ma  ¹✉ & Junping Gao  ¹✉

Modern rose (*Rosa hybrida*) is a recently formed interspecific hybrid and has become one of the most important and widely cultivated ornamentals. Here we report the haplotype-resolved chromosome-scale genome assembly of the tetraploid *R. hybrida* 'Samantha' ('JACmantha') and a genome variation map of 233 *Rosa* accessions involving various wild species, and old and modern cultivars. Homologous chromosomes of 'Samantha' exhibit frequent homoeologous exchanges. Population genomic and genomic composition analyses reveal the contributions of wild *Rosa* species to modern roses and highlight that *R. odorata* and its derived cultivars are important contributors to modern roses, much like *R. chinensis* 'Old Blush'. Furthermore, selective sweeps during modern rose breeding associated with major agronomic traits, including continuous and recurrent flowering, double flower, flower senescence and disease resistance, are identified. This study provides insights into the genetic basis of modern rose origin and breeding history, and offers unprecedented genomic resources for rose improvement.

Rose is among the most important ornamental plants and provides a precious source of natural scent, generating billions of dollars in trade worldwide annually¹. It has been cultivated for thousands of years in China and Europe², separately but coincidentally. The ancient rose cultivars, or so-called old roses, originated from reticulate evolution among several wild species within the genus *Rosa* of different ploidy levels³. As ancestors of modern rose, the wild species that have been passed down for thousands of years exhibit important agronomic/ornamental traits, including double flower, continuous and recurrent flowering, unique scent, flower shape, inflorescence type, growth vigour and abiotic/biotic stress tolerance^{4,5}. These wild species exhibit a broad geographic distribution, with *R. odorata* var. *gigantea*, *R. chinensis* var. *spontanea*,

R. chinensis 'Old Blush', *R. wichuraiana*, *R. rugosa* and *R. fedtschenkoana* predominantly found in China, *R. moschata* in South and West Asia, and *R. gallica* and *R. canina* throughout Europe².

Since the eighteenth century, several diploid Chinese old cultivars including *R. chinensis* 'Slater's Crimson China', *R. chinensis* 'Old Blush', *R. odorata* 'Hume's Blush Tea-scented China' and *R. odorata* 'Park's Yellow Tea-scented China' had been brought to Europe where they were crossed with European old cultivars and other wild species⁴, producing a series of intermediate types that include Bourbon, Noisette, Portland, Tea, Hybrid Perpetual and so on. In 1867, the hybridization effort met with great success when a hybrid named 'La France' was created by the rosarian Jean-Baptiste André Guillot. This hybrid rose, which combined

A full list of affiliations appears at the end of the paper. ✉ e-mail: zf25@cornell.edu; ma_nan@cau.edu.cn; gaojp@cau.edu.cn

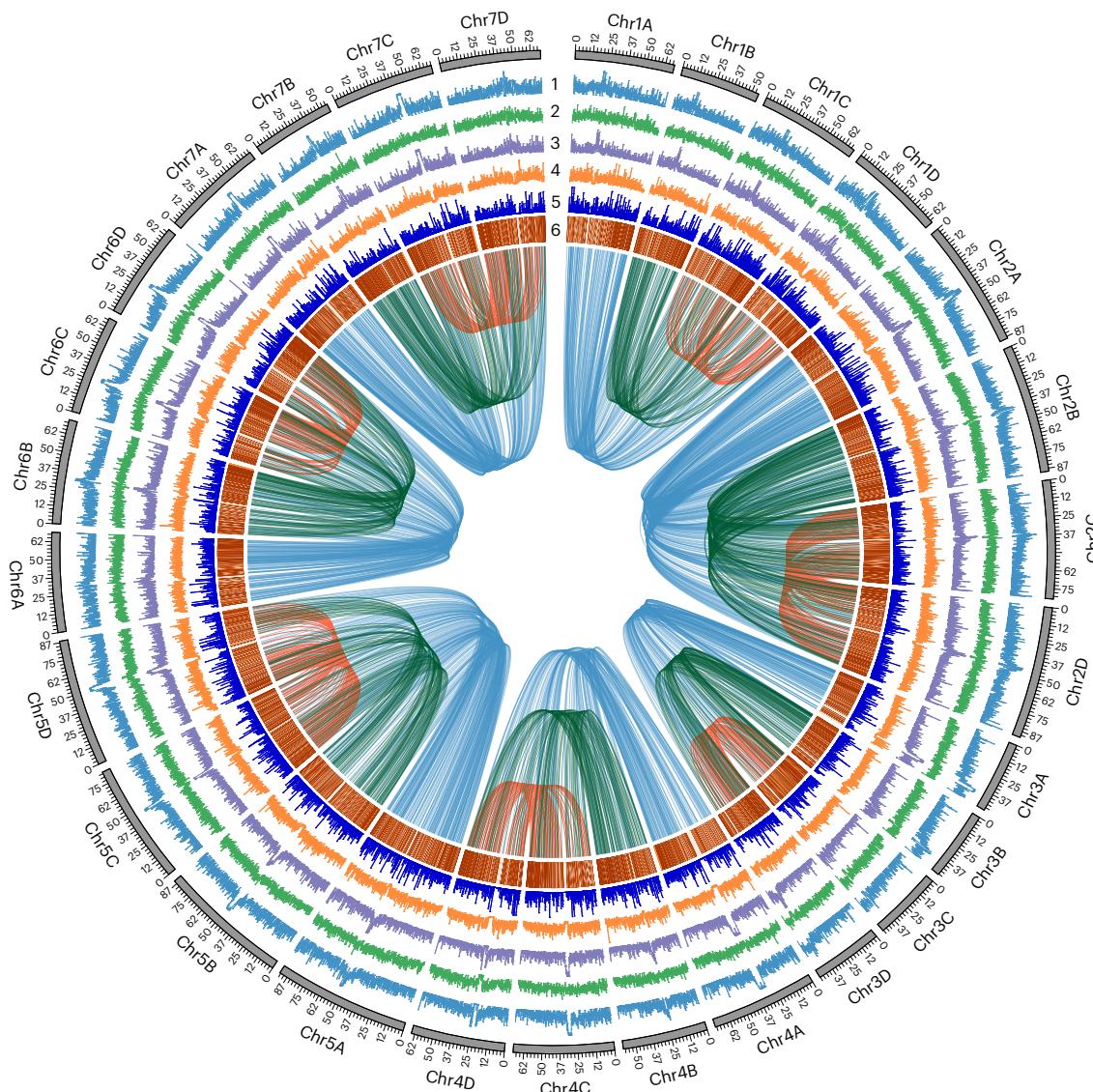


Fig. 1 | Genomic landscape of *R. hybrida* 'Samantha'. 1–6, LTR-RT density (1), Copia-type LTR-RT density (2), Gypsy-type LTR-RT density (3), DNA transposon density (4), gene density (5) and gene expression levels (6) in 500 kb non-overlapping windows. Syntenic blocks between homologous chromosomes are shown at the centre.

the good traits of Chinese and European old roses, has since become recognized as the class of modern rose (referred to as *Rosa hybrida*). Modern rose shows great phenotypic variation, as well as genetic potential for yield, quality and phenological adaptability, making it quickly predominate in rose cultivation^{5,6}. Nowadays, over 35,000 modern rose cultivars, most of which are tetraploid, have been cultivated worldwide. These modern roses have a wide variety of colours, flower shapes and scent, as well as different plant architectures, all of which are important traits and need to be better understood.

According to documentary records, at least 8–20 wild species and old cultivars with different ploidy levels were involved in the creation of modern roses⁴. The high heterozygosity resulting from reticulate evolution, high percentage of repetitive sequences and polyploidy all challenge deciphering the mystery of the genomic composition of modern roses. The genome of a doubled haploid (DH) line of *R. chinensis* 'Old Blush', which has been considered as one of the important progenitors of modern rose, has become available recently^{7,8}. However, genomic information of tetraploid modern roses is scarce, and their origin, evolution and improvement remain to be investigated. Here we present (1) a haplotype-resolved high-quality genome assembly of a tetraploid modern rose cultivar 'Samantha' ('JACmantha'), which is

characterized by its red flowers, mild scent, medium-sized high-centred blooms, continuous and recurrent flowering traits, and leathery foliage (Extended Data Fig. 1a) and (2) genome resequencing of 233 rose accessions involving wild species, and old and modern rose cultivars. We present genome features of a modern rose cultivar, characterize the ancestral contributions of initial founder species and identify selective sweeps during modern rose breeding history that are associated with important traits. Altogether, our work sheds light on the complex ancestry of modern roses and how human selection has reshaped rose genomes in a short period of less than 200 years. Our results further provide valuable information for understanding the genetic basis of ornamental and cultivated traits in roses and will facilitate future breeding of this important ornamental crop.

Results

Haplotype-resolved genome assembly of tetraploid modern rose

To develop a reference genome for tetraploid modern roses, we sequenced the genome of cultivar 'Samantha' ($2n = 4x = 28$) (Extended Data Fig. 1b), which has an estimated tetraploid genome size of ~1.91–2.13 Gb on the basis of k -mer and flow cytometry analyses

(Supplementary Figs. 1 and 2, and Table 1). We generated 103 Gb of PacBio HiFi reads, 237 Gb of Oxford Nanopore Technologies (ONT) ultra-long reads, 140 Gb of MGI paired-end reads, 543 Gb of high-throughput chromatin conformation capture (Hi-C) and 43 Gb of Pore-C⁹ sequences (Supplementary Table 2). The final genome assembly contained 856 contigs, with a total size of 2,151 Mb and an N50 length (the length such that half of all sequence is in contigs of this size or larger) of 37.76 Mb. The Pore-C data were used to assemble 91.6% of the contigs (1,971 Mb) into 28 pseudochromosomes, which were further manually corrected on the basis of Hi-C reads mapping (Fig. 1 and Supplementary Table 3). The homologous chromosomes of 'Samantha' were named on the basis of their potential relative relationship with 17 wild species distributed in China or Chinese old roses (Supplementary Note 1).

To evaluate the quality of the assembled 'Samantha' genome, we first aligned the MGI paired-end and HiFi reads to the genome, resulting in high mapping rates of 99.76% and 99.95%, respectively. In addition, ~96.31–99.12% of the RNA-seq reads from different tissues could be aligned to the assembly (Supplementary Table 2). Furthermore, BUSCO¹⁰ evaluation showed that 98.7% of the core genes were completely captured in the 28 'Samantha' chromosomes, with the majority (84.2%) present in four copies (Supplementary Table 4 and Supplementary Fig. 3). The LTR assembly index¹¹ (LAI) for the 'Samantha' genome was high (21.93). Similarly, *k*-mer analysis demonstrated the high completeness of the genome assembly (Supplementary Fig. 4), with a consensus quality value and *k*-mer completeness rate of 53 and 97.58%, respectively. The integrity of the chromosome-scale assembly of 'Samantha' was supported by the K5 genetic map of tetraploid modern rose¹² (Supplementary Fig. 5) and the high degree of synteny with the genome of *R. chinensis* 'Old Blush'⁸ (Extended Data Fig. 2). Both Pore-C and Hi-C signal heat maps supported the high continuity and phasing accuracy of the assembly (Extended Data Fig. 3). The phasing quality of the 'Samantha' genome was further evaluated using nPhase¹³, which indicated a very low level of potential phase switching errors (0.07%) in the assembly (Supplementary Fig. 6). Together, these results confirmed the high accuracy and completeness of the phased chromosome-level assembly of the 'Samantha' genome. Finally, the quality of the assembled genome of 'Samantha' is better than or comparable to the published genomes from the genus *Rosa* (Supplementary Table 5).

A total of 59.32% of the 'Samantha' genome assembly was identified as repeat sequences, with the majority (39.02% of the assembly) annotated as long terminal repeat retrotransposons (LTR-RTs) (Supplementary Table 6). A total of 141,827 protein-coding genes were predicted from the genome, of which 130,192 (91.79%) were functionally annotated (Supplementary Table 7).

Genomic variation map

We performed genome resequencing of 215 accessions in the genus *Rosa*, including wild accessions from different sections of the genus (Supplementary Notes 2 and 3), intermediate old cultivars (bred from hybridization among old cultivars from China and Europe) and

modern cultivars (Supplementary Table 8). Along with 18 accessions with publicly available resequencing data⁸, a total of 233 accessions were analysed in this study, including 62 (wild species and old cultivars) from section Chinenses, 16 from section Rosa, 18 from section Synstylae, 11 from section Cinnamomeae, 4 from section Caninae, 15 from other sections (including Pimpinellifoliae, Microphyllae, Banksiae, Bracteatae and Laevigatae), 65 intermediate old cultivars and 42 modern cultivars. In total, we identified 17,567,247 single nucleotide polymorphisms (SNPs), among which 12,958,041 (73.76%) were identified within Chinenses, 11,359,119 (64.66%) within Rosa and 12,396,729 (70.57%) within Synstylae. As for genomic locations, 7,297,927 (41.54%) were in intergenic regions, 3,837,431 (21.84%) in intron regions and 1,917,739 (10.92%) in coding regions. Among SNPs in coding regions, 999,674 (52.13%) were non-synonymous and 893,619 (46.60%) were synonymous (Supplementary Table 9). Furthermore, a total of 21,826 SNPs introduced premature stop codons in 14,873 genes. Among these nonsense SNPs, we identified several SNPs affecting key genes involved in regulating various plant growth and development pathways (Supplementary Table 10), providing a valuable resource for future studies on rose plant growth and development.

We performed kinship analysis for the 233 rose accessions using the KING programme¹⁴ and did not identify any duplicated accessions or first-degree relatives in this rose collection. We then evaluated the nucleotide diversity of different rose populations. The nucleotide diversity (π) of six different groups (sections Chinenses, Rosa, Synstylae, Cinnamomeae, Caninae and other sections) ranged from 1.42×10^{-3} to 4.11×10^{-3} (Supplementary Table 11). The intermediate old cultivars showed the highest degree of nucleotide diversity (5.24×10^{-3}), also with 99.45% of the SNPs being detected in this group. The nucleotide diversity of modern cultivars displayed a slightly decreased level (4.79×10^{-3}) compared with that of the intermediate old cultivars, but notably higher than any other groups (Supplementary Table 11). This suggests a weak bottleneck effect during the improvement of modern roses and indicates that modern roses still have great genetic potential for further enhancement.

Origin of modern rose

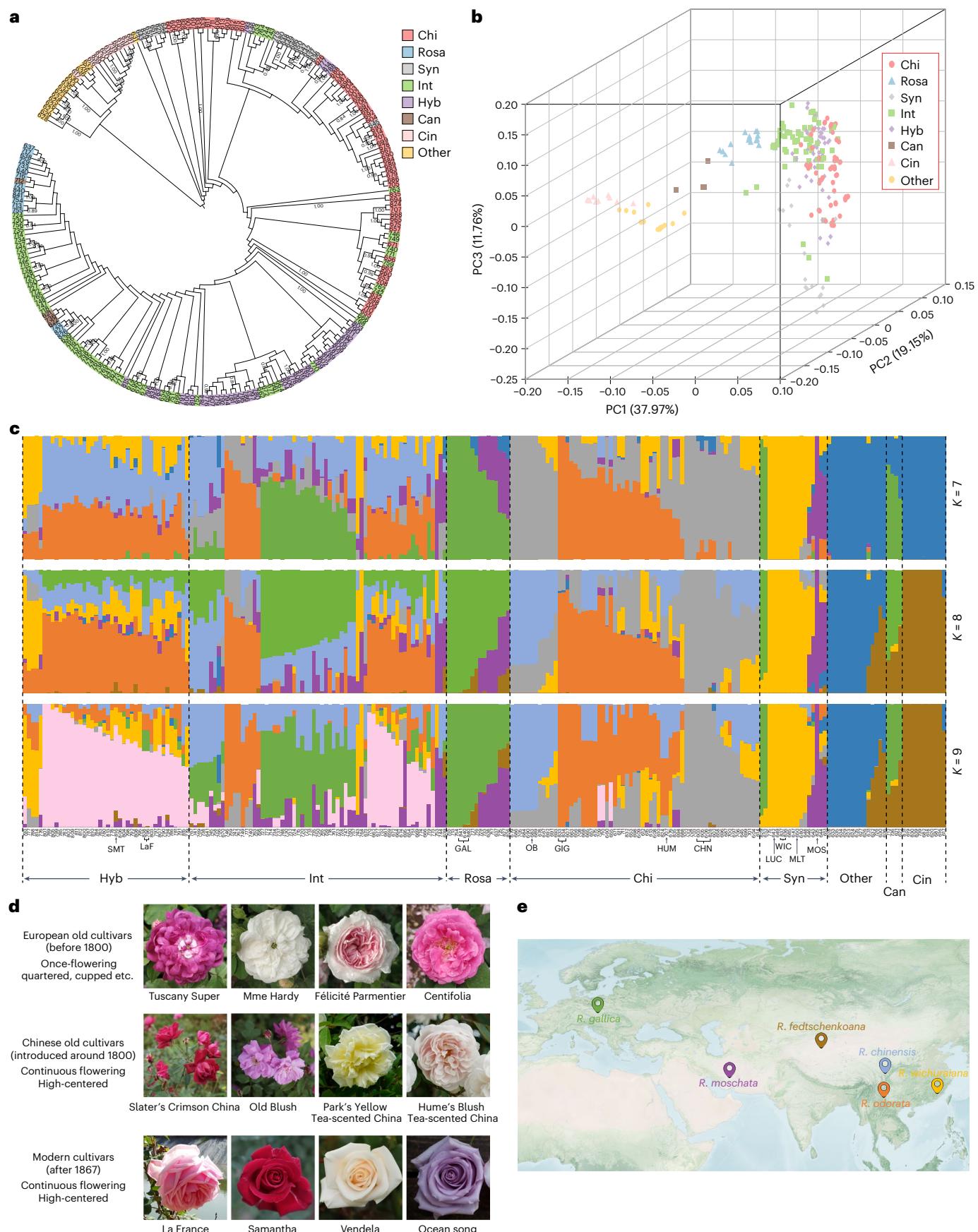
According to historic documents and genetic analysis, 8–20 species of *Rosa* are thought to be involved in modern rose creation⁵, although the exact donors are still in dispute. To better understand the origin and breeding history of modern rose as a new population, a phylogenetic tree of the 233 accessions was constructed using high-quality SNPs (Fig. 2a). Modern cultivars were located within the clade comprising all members of section Chinenses and part of section Synstylae, suggesting possible major contributions of sections Chinenses and Synstylae species to modern cultivars. Intermediate old cultivars, as early hybrids, were mainly distributed around modern cultivars on the phylogenetic tree. In the principal component analysis (PCA), accessions from section Cinnamomeae and other sections were consistently clearly separated from modern cultivars (Fig. 2b). The intermediate group exhibited the broadest diversity. By contrast, modern cultivars formed a tighter cluster within the intermediate group, indicating a narrower genetic

Fig. 2 | Population structure of 233 accessions within the genus *Rosa*.

a, Phylogenetic tree constructed using SNPs at 4-fold synonymous third-codon transversion (4DTv) sites. Hyb, modern cultivars, *R. hybrida*; Int, intermediate old cultivars; Syn, section Synstylae; Chi, section Chinenses; Rosa, section Rosa; Can, section Caninae; Cin, section Cinnamomeae; Other, sections Pimpinellifoliae, Microphyllae, Bracteatae, Banksiae, and Laevigatae. **b**, Principal component analysis of the 233 accessions. **c**, Bayesian model-based clustering of the 233 accessions with the number of ancestry kinships (K) from 7 to 9. Each vertical bar represents an accession. Each colour represents one putative ancestral background and the y axis quantifies ancestry contribution. SMT, *R. hybrida* 'Samantha'; LaF, first Hybrid Tea rose cultivar 'La France'; HUM, *R. odorata* 'Hume's Blush Tea-scented China'; OB, *R. chinensis* 'Old Blush'; WIC,

R. wichuriana; MLT, *R. multiflora* 'Thornless'; LUC, *R. luciae*; MOS, *R. moschata*; GIG, *R. odorata* var. *gigantea*; CHN, *R. chinensis* var. *spontanea*; GAL, *R. gallica*. **d**, Changes in flower shape along the rose breeding history.

e, Geographical distribution of the six original species. *R. chinensis* is distributed in Sichuan, Hubei and Guizhou in China. *R. wichuriana* is distributed in Zhejiang, Guangdong, Guangxi, Fujian and Taiwan in China. *R. gallica* is distributed in central and southern Europe and western Asia. *R. fedtschenkoana* is distributed in Xinjiang in China, Central Asia, the Tianshan Mountains and the Pamirs. *R. odorata* is distributed in Yunnan in China. *R. moschata* is believed to have originated from South and West Asia, including Afghanistan, Pakistan, India and Iran. The geographic map used here was obtained from NASA (<https://visibleearth.nasa.gov/images/147190/explorer-base-map/147191w/>).



variation after human selection along the continuous breeding history. The closer distribution of modern cultivars to Chinese accessions (section Chinenses and part of section Synstylae) was consistent with the increased Chinese/European allele ratio discovered in cultivars bred in the eighteenth to nineteenth century¹⁵.

To infer possible progenitors of modern roses, we further performed population structure analysis (Fig. 2c). Cross-validation error analysis revealed that seven populations ($K = 7$) were optimal for these 233 accessions (Supplementary Fig. 7). At $K = 7$, accessions from section Rosa were quite distinguishable from other accessions and they genetically contributed to part of intermediate old cultivars rather than modern cultivars. The section Chinenses was clearly divided into two groups, one consisting of *R. chinensis* and the other consisting of *R. odorata*. At $K = 8$, *R. chinensis* 'Old Blush' was separated from the original wild species *R. chinensis* var. *spontanea* and was quite genetically different from 'Hume's Blush Tea-scented China', which was derived from wild species *R. odorata* var. *gigantea*. For modern cultivars, they mainly harboured genetic compositions from sections Chinenses, Rosa and Synstylae. We found that, 'Hume's Blush Tea-scented China', a hybrid cultivar derived from *R. odorata* var. *gigantea* and *R. chinensis*, contributed more to the genetic composition of modern cultivars than *R. chinensis* 'Old Blush'. At $K = 9$, modern roses showed similar genetic components among each other even with complex origins. Considering that favoured mutations can be selected and well preserved through clonal propagation in roses, it is not surprising that modern roses have developed from several heterogeneous ancestors into a genetically relatively similar group.

Within modern roses, cultivars with larger genetic contributions from section Synstylae (Fig. 2c) have flowers in large clusters or trusses and were thus classified as group Floribunda (Supplementary Note 3). It has long been recognized that genome introgression from *R. multiflora* into modern cultivars has passed down its typical inflorescence trait (flowers in clusters) to Floribunda, and this is supported by morphological evidence (Extended Data Fig. 4). Genetic structure of all the three wild species from section Synstylae, *R. multiflora*, *R. luciae* and *R. wichuraiana*, could not be further distinguished from each other (Fig. 2c), hence the exact contribution of each species to this inflorescence trait of modern roses still remains unclear.

Before the introduction of Chinese germplasms around 1800, European old cultivars (for example, Gallica, Damask, Alba and Centifolia; Fig. 2d, top) were generally once flowering and had a relatively similar quartered or cupped flower shape. Relatively detailed records of introduced Chinese accessions focused on the four well-known cultivars (Fig. 2d, middle). Continuous and recurrent flowering is a highly valued ornamental trait and *R. chinensis* 'Old Blush' exhibiting this feature is always believed to be an important Chinese progenitor to modern roses^{2,7,8}. Unexpectedly, its contribution to modern roses at the genomic level seemed not large (Fig. 2c). In addition to the continuous and recurrent flowering trait of the four introduced Chinese old cultivars, the dark red colour of *R. chinensis* 'Slater's Crimson China' and the distinct tea scent from two other *R. odorata* cultivars ('Hume's Blush Tea-scented China' and 'Park's Yellow Tea-scented China') made them all extensively used in hybridizations for modern rose formation. The two *R. odorata*-derived cultivars are believed to only contribute to the tea-scent trait, which has been lost in most modern rose cultivars. Interestingly, our population structure analysis suggested that they contributed the most to the genetic composition of modern cultivars ($K = 7$ –8 in Fig. 2c) and may thus have laid the foundation of the favoured flower shape of modern cultivars (Fig. 2d, bottom). It is worth mentioning that *R. odorata*-derived old cultivars exhibit a high-centred flower shape distinct from that of European old cultivars, and their petal edges can be reflexed backward at later opening stages. The flower shape of modern rose cultivars (Fig. 2d, bottom) mimics that of *R. odorata*-derived old cultivars, with increased petal numbers. Collectively, *R. odorata* and derived cultivars contributed a larger portion

to the genome structure and favoured flower shape of modern cultivars than *R. chinensis* 'Old Blush' and *R. gallica* (Fig. 2c,d). Ultimately, we clarified the contribution of six potential original species to modern cultivars, and the geographical distribution¹⁵ of these species highlights the importance of artificial hybrid breeding (Fig. 2e).

Genomic composition of *R. hybrida* 'Samantha'

The complex hybrid history of modern roses makes it difficult to identify the origins of their chromosomes. Population structure analysis reveals six potential original species of modern roses (Fig. 2c). To better analyse the genomic composition of *R. hybrida* 'Samantha', the origins of different regions of each chromosome were inferred on the basis of the mapped read depth and the variant frequency of the original wild species. The consensus results of the two approaches were used to infer possible ancestral origins of regions in each chromosome (Fig. 3a,b and Extended Data Fig. 5). We inferred that the largest proportion of 'Samantha' genomes is derived from the *R. odorata* var. *gigantea*, accounting for 18.25%, and *R. chinensis* 'Old Blush', *R. wichuraiana*, *R. gallica*, *R. moschata* and *R. fedtschenkoana* contribute 17.51%, 13.47%, 11.80%, 10.79% and 7.73% of the genome, respectively (Fig. 3c). Interestingly, the 'Samantha' chromosomes exhibited a high degree of chimaerism, each derived from multiple ancestral species except Chr7B, indicating extensive homoeologous exchanges and/or introgressions in the 'Samantha' genome (Extended Data Fig. 5) and consistent with the segmental allotetraploid nature of modern rose¹². Chr7B was derived from *R. fedtschenkoana* (section Cinnamomeae) and was notably shorter than the other three homologous chromosomes, which we speculate might be the reason for the absence of homoeologous exchanges. Furthermore, a large inversion between the homologous chromosomes was observed on chromosome 7 (Fig. 3d), which was further supported by the local Pore-C heat maps (Extended Data Fig. 6) and the collinearity with the four published *Rosa* genomes^{8,16–18} (Extended Data Fig. 7). Genomic regions of Chr7A and Chr7C harbouring this inversion were both derived from *R. odorata* var. *gigantea*, while those of Chr7B and Chr7D were derived from *R. fedtschenkoana* and *R. wichuraiana*, respectively.

Artificial selection during modern rose breeding

To investigate genomic regions that have been under selection during rose breeding, we conducted selective sweep analysis between modern cultivars (cultivars bred after 1867; $n = 32$) and intermediate old cultivars (cultivars bred before 1867; $n = 45$). Considering that the ploidy level could interfere with this analysis, we chose cultivars confirmed to be tetraploid from intermediate old cultivars (referred to as Int_4) and modern cultivars (referred as Hyb_4) for this analysis (Supplementary Table 8). Int_4 had a higher nucleotide diversity ($\pi = 5.39 \times 10^{-3}$) than Hyb_4 ($\pi = 4.83 \times 10^{-3}$), reflecting the function of human selection during cultivar improvement (Fig. 4a). In addition, Int_4 exhibited a faster linkage disequilibrium decay than Hyb_4 (Fig. 4b), again supporting the decreased genetic diversity for recently formed modern cultivars. By comparing Int_4 with Hyb_4 and using combined filtering of π ratio and fixation index (F_{ST}), selective sweep regions were identified with a total length of 44.42 Mb, which contained 3,382 genes (Supplementary Tables 12 and 13). Within these regions, genes related to continuous and recurrent flowering, flower meristem and floral organ development, flower colour, senescence, growth, disease resistance and prickle formation were revealed (Fig. 4c).

The ability for continuous and recurrent flowering is among the most important traits of modern roses. It is the basis for cut rose production and is also highly valued for garden/potted roses. This character is believed to be inherited from *R. chinensis*^{2,8}. The *KSN* gene in rose has been found to be involved in the determination of once/continuous flowering^{19,20}. Here, the *KSN* genes (*Rh3AG182500*, *Rh3BG208400*) were found in a selective sweep region (Extended Data Fig. 8a,b,d,e). Clustering analysis of SNP genotype in this selective sweep region

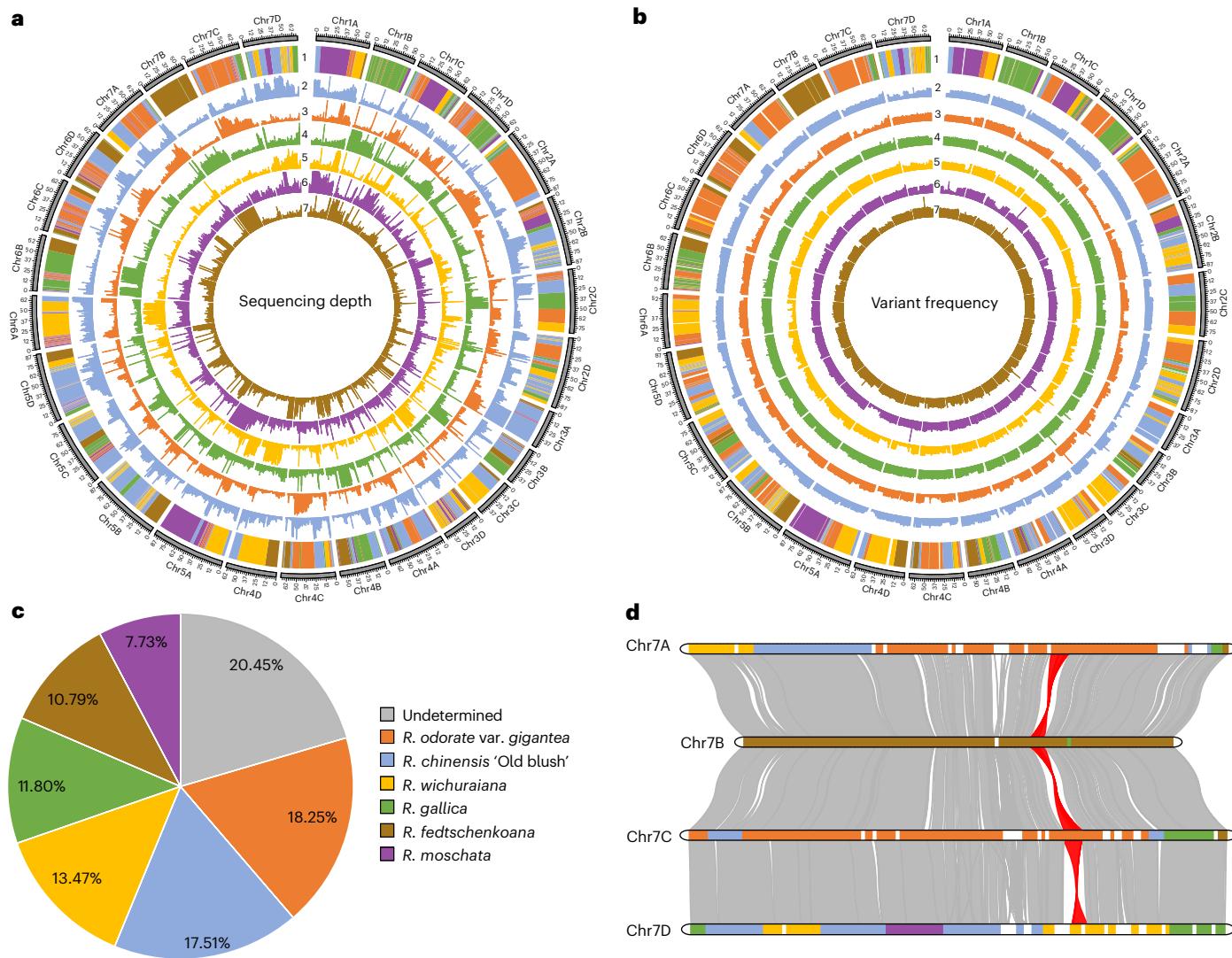


Fig. 3 | Genetic compositions of the 'Samantha' chromosomes. **a**, Origins of different regions on the chromosomes (1) inferred by the depth coverage of reads from wild ancestors *R. chinensis* 'Old Blush' (2), *R. odorata* var. *gigantea* (3), *R. gallica* (4), *R. wichuraiana* (5), *R. moschata* (6) and *R. fedtschenkoana* (7). **b**, Origins of different regions on the chromosomes (1) inferred by the variant frequency between 'Samantha' and wild ancestors *R. chinensis* 'Old Blush' (2), *R. odorata* var. *gigantea* (3), *R. gallica* (4), *R. wichuraiana* (5), *R. moschata* (6)

and *R. fedtschenkoana* (7). **c**, Summary statistics of contributions of the six original species to the 'Samantha' genome. **d**, Collinearity of the homologous chromosomes of chromosome 7. Colours on the chromosomes represent origins from different original species *R. chinensis* 'Old Blush' (blue), *R. odorata* var. *gigantea* (orange), *R. gallica* (green), *R. wichuraiana* (yellow), *R. fedtschenkoana* (brown) and *R. moschata* (purple). The inversion is shown in red.

further supported the contribution of *R. chinensis* to the continuous and recurrent flowering of modern roses (Extended Data Fig. 8c,f). Furthermore, genes encoding regulators involved in vernalization²¹ (for example, *VIL1*, *VIP*, *FRL*, *RTV1*) and the photoperiod pathway^{22,23} (for example, *NFL*, *LATE*) were also intensively selected, suggesting that the trait of continuous and recurrent flowering may have been realized by fixing genes in separate flowering pathways.

Flower meristem identity genes *SPL4* (ref. 24), *CLV1* (ref. 25), *BAM1* (ref. 26) and *LMI1* (ref. 27) were found to have undergone selection during modern rose breeding. They participate in the formation of the shoot apical meristem and affect the development of floral organs. The number of petals is another important ornamental trait and has been increased greatly during modern rose breeding. In rose, the floral identity gene *AP2* has been proven to regulate petal numbers through stamen petaloidy^{28–30}. Here we identified that *AP2* genes (*Rh2AG184600*, *Rh2BG195000*) have been under selection during modern rose breeding (Extended Data Fig. 9a,b,d,e). We performed

clustering on the SNP genotypes of these two selective sweep regions that contained *AP2* genes and found that SNP profiles of modern cultivars were highly similar to those of *R. odorata* (Extended Data Fig. 9c,f). This directly demonstrates the crucial contribution of *R. odorata* to the petal number in modern roses. Moreover, flower colour is among the most attractive characteristics of modern roses. *GT5* involved in floral pigment biosynthesis³¹ was also identified in selective sweep regions.

Opening and senescence of flowers are essential parameters determining the quality of modern roses. Ethylene has been extensively documented to have a pivotal role in governing flower opening and senescence in rose^{32–34}. Ethylene-related genes were found in several highly selected regions, including biosynthesis-related genes (*ACOs*) and genes encoding signalling components (*CTR1*, *EIN2*) and ethylene-responsive transcription factors (ERFs), indicating that both ethylene biosynthesis and signalling pathways have been selected during the breeding history of modern roses. *CTR1* (ref. 35) (*Rh3BG199500*), the key component of the ethylene signal transduction pathway, has

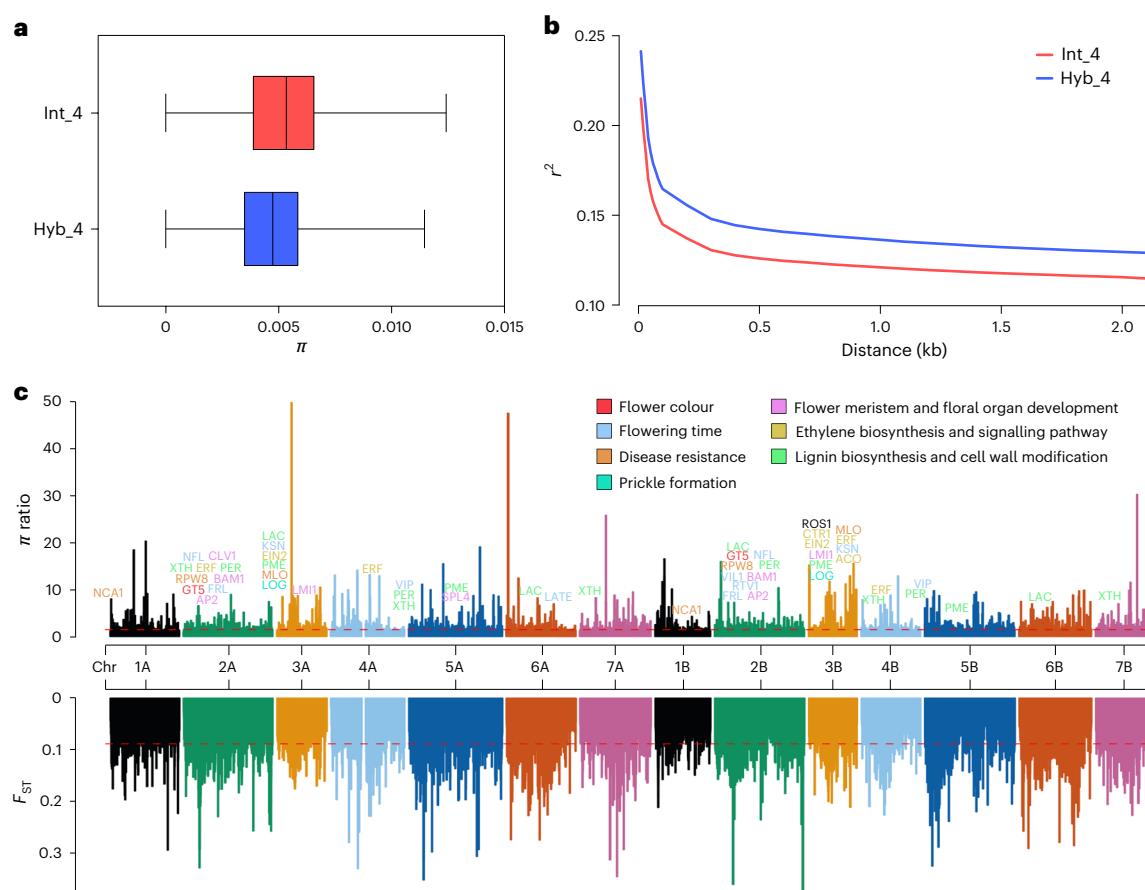


Fig. 4 | Genome-wide distribution of selective sweeps in *R. hybrida*. **a**, Nucleotide diversity (π) of tetraploid intermediate old cultivars (Int_4; $n = 45$) and tetraploid modern rose cultivars (Hyb_4; $n = 32$). For boxplots, the lower and upper bounds of the box represent the 25th and 75th percentiles, respectively; the centre line marks the median; and the whiskers extend to the full range of the data. P -value (1.28×10^{-263}) was calculated using a two-sided Student's t -test. **b**, Linkage disequilibrium decay in Int_4 and Hyb_4. **c**, π ratio and F_{ST} analyses between Int_4 and Hyb_4 reveal potential selective regions across rose chromosomes. Horizontal red dashed lines indicate the top 5% cut-off values. Gene abbreviations: NCA1, NO CATALASE ACTIVITY1; KSN,

Koushin; CLV1, CLAVATA1; NFL, NO FLOWERING IN SHORT DAY; FRL, FRIGIDA-like; VIL1, VERNALIZATION INSENSITIVE3-like1; SPL4, squamosa promoter binding protein-like4; LMI1, LATE MERISTEM IDENTITY1; LATE, LATE FLOWERING; BAM1, BARELY ANY MERISTEM1; VIP, VERNALIZATION INDEPENDENCE; GTS, UDP-glucose flavonoid 3-O-glucosyltransferase5; RTV1, RELATED TO VERNALIZATION1; AP2, APETALA2; CTR1, CONSTITUTIVE TRIPLE RESPONSE1; ACO, 1-aminocyclopropane-1-carboxylate oxidase; EIN2, ETHYLENE INSENSITIVE2; ERF, ethylene response factor; LAC, LACCASE; PER, PEROXIDASE; PME, pectinesterase; XTH, xyloglucan endotransglucosylase/hydrolase; RPW8, resistance to powdery mildew8; MLO, mildew resistance locus O; ROS1, REPRESSOR OF SILENCING1; LOG, LONELY GUY.

been under selection, and its SNP genotype profile was similar to that of *R. chinensis* (Extended Data Fig. 10a,b). Furthermore, *ROS1* (ref. 36) (*Rh3BG199400*), encoding a DNA demethylase, was also found in this selective sweep region (Extended Data Fig. 10c), indicating that DNA methylation might have a crucial role in determining rose traits during its breeding. However, the current results cannot rule out the possibility that the selection of *ROS1* may be due to a hitchhiking effect from the selection of *CTR1*, or vice versa.

In addition, genes encoding enzymes related to lignin biosynthesis and cell wall modification (for example, XTH³⁷, LAC, PER and PME) were found in selective sweep regions. Mutations in LAC and PER could result in defects in lignification and thus growth arrest^{38,39}. PMEs were found to be associated with plant height in soybean⁴⁰ and fruit weight in peach⁴¹. Therefore, selection of these genes in modern roses might be related to growth rate and yield. In modern rose breeding, selection for disease resistance has always been a main target⁴². Therefore, it is not surprising that compared with intermediate old cultivars, disease resistance in modern cultivars has been selected during the past 200 years of modern rose breeding history. In the selective sweep regions, genes homologous to *Arabidopsis* *NCA1* (ref. 43) that have a crucial role in the clearance of hydrogen peroxide—*Arabidopsis* *RPW8*

(ref. 44) involved in powdery mildew resistance, and rose *MLO* gene⁴⁵, also responsible for powdery mildew resistance—were identified. Plant NB-LRR immune receptors have evolved the ability to specifically recognize a wide range of effector proteins from different pathogens⁴⁶. Within the selective sweep regions, we identified 14 TIR-NBARC-LRR (TNL) and 12 CC-NBARC-LRR (CNL) immune receptors (Supplementary Table 14). Selection of these genes further confirms the importance of disease resistance in breeding of modern cultivars. A recent study revealed that the *LOG* gene directly regulates prickle formation in rose⁴⁷. This gene was identified in selective sweep regions, indicating that the prickle trait has been influenced by artificial selection during modern rose breeding.

The absence of a rich scent in a large number of modern cultivars is an unexpected trait overlooked during breeding. Our study did not identify any genes related to scent synthesis in genome regions under selection. A previous study has shown that the tea-scent trait of modern roses originates from Chinese roses, with the *OOMT* genes playing a key role⁴⁸. We were able to identify the *OOMT1* gene (*Rh2AG330600*) located in a region on chromosome Chr2A of the 'Samantha' genome that is derived from *R. odorata* var. *gigantea*. No *OOMT* genes were identified in other homologous chromosomes of 'Samantha'. While

the *OOMT1* gene was not identified to be under selection, an SNP (Chr2A:47717127) in this gene was identified in our rose population, corresponding to the critical mutation site of *OOMT1* where the presence of adenine (A) at this locus leads to the development of the tea-scent trait in roses⁴⁸. Analysis of the genotypes of this locus in the 233 rose accessions further confirmed that the tea-scent trait originates from the section Chinese (Supplementary Table 15). Given that the breeding of modern roses have only occurred for a few hundred years, we believe that scent could become a new targeted trait in future rose breeding to meet consumer preference, and our genotype analysis of the *OOMT1* gene provides useful information that will facilitate this process.

Discussion

The modern rose, which is mostly tetraploid, is a hybrid resulting from the crossbreeding of multiple ancestors from both east and west of Eurasia. It has inherited important agronomic traits from its European and Chinese progenitors. While one of the Chinese diploid progenitor *R. chinensis* 'Old Blush' has been sequenced recently^{7,8}, a reference genome of the tetraploid modern rose is of irreplaceable importance for understanding its origin and breeding, as well as for improving this important ornamental crop. In this study, we present a haplotype-resolved genome assembly of the tetraploid modern rose cultivar 'Samantha', along with a comprehensive genome variation map of 233 accessions from wild species and modern and old cultivars in the genus *Rosa*. This work allows us to reconstruct the history of divergence, hybridization and breeding of roses. The genome of 'Samantha' is one of the few phased chromosome-scale polyploid plant genome assemblies reported so far, providing a powerful resource for future comparative and evolutionary genomic studies.

Population genomic analyses reveal that Chinese and European old cultivars originated from reticulate evolution among several wild species within *Rosa*. *R. odorata*, in addition to *R. chinensis* 'Old Blush'^{4,15}, played an important role as the Chinese progenitor of modern rose during its creation through hybridizations in the nineteenth century. Notably, some previously unappreciated wild species, such as those in section Synstylae, have also played important roles in the formation of modern rose, particularly the Floribunda type with clustered flowers. In the population structure analysis, it is observable that the genetic background of *R. chinensis* 'Old Blush' has substantially infiltrated the intermediate old cultivars. This is consistent with the rose breeding history, demonstrating the key role of *R. chinensis* 'Old Blush' in the formation of the intermediate group. Meanwhile, the modern cultivars have the largest proportion of the genetic background corresponding to *R. odorata*, suggesting that *R. odorata* may have been utilized by breeders later compared with *R. chinensis* 'Old Blush'.

Over thousands of years, the cultivation and domestication of old cultivars in both China and Europe have accumulated a substantial number of ideal agronomic traits. These traits have been retained and further selected in modern rose, making it one of the most important ornamental crops worldwide. To identify genome regions under selection during modern rose breeding that are potentially associated with important agronomic traits (such as continuous and recurrent flowering, scent, senescence and disease resistance), we compared modern roses with old cultivars. Continuous and recurrent flowering is a core trait of modern rose cultivars, and therefore selection of related genes is expected. On the other hand, extending the ornamental period of modern roses is important, especially for cultivars used for cut flowers. It is important to develop cultivars that have low sensitivity to ethylene to delay senescence. By controlling the crucial component upstream of the ethylene signalling cascade, it is possible to effectively reduce ethylene signal transduction, thereby minimizing the ethylene response and extending the life of cut flowers. It has been reported that transposon insertion in the *AP2L* gene causes variation in its coding region, which makes it resistant to miRNA172 cleavage, promoting the double flower phenotype²⁸. Moreover, transposon insertion in the *KSN*

gene restricts its expression, leading to the continuous and recurrent flowering trait of modern roses²⁰. In 'Samantha', both *AP2L* and *KSN* harbour the corresponding transposon insertions, and the selection of *AP2L* and *KSN* during modern rose breeding may have played a key role in the development of the double flower and continuous and recurrent flowering traits observed in modern roses.

It is worth noting that in this study, allele dosage was not considered in the selective sweep analysis due to the lack of a theoretical model/software specifically developed for this purpose. Although this may result in missing some interesting genomic regions, for example, those that have been selected during breeding on the basis of their dosages, our selective sweep analysis still yields abundant novel and valuable information that is pertinent to the genetic history of modern rose breeding. Furthermore, our comprehensive genomic and population genomic analyses provide substantial insights into the structure and evolution of the tetraploid modern rose genome, nucleotide diversity and fixation of horticultural traits. Together with the haplotype-resolved high-quality modern rose genome and single-base-resolution variation map, the findings of this study provide valuable information for future research on the agronomically important traits of roses and facilitate marker-assisted and genomic selection-based breeding in the future.

Methods

Sample preparation and sequencing

The widely cultured modern rose cultivar 'Samantha' was selected for genome sequencing. High-quality genomic DNA was extracted using the cetyltrimethylammonium bromide⁴⁹ (CTAB) method. The ONT ultra-long library was constructed using the Nanopore Ligation Sequencing kit (SQK-LSK114) and sequenced on the PromethION (ONT). PacBio HiFi libraries with inserts of ~20 kb were prepared following the PacBio HiFi library construction protocol and sequenced on the PacBio Revio system (Pacific Biosciences). For MGI sequencing, a DNA sequencing library was constructed with an insert size of ~300 bp and sequenced on a DNBSEQ-T7 platform (BGI). The Pore-C libraries were generated according to the method described in ref. 50 with improvements. Briefly, ~0.25 g of young fresh leaves were ground to a fine powder in liquid nitrogen and transferred to 1 ml of pre-cold extraction buffer (0.4 M sucrose, 10 mM Tris-HCl, 10 mM MgCl₂, 0.2% Triton X-100, 1 mM EDTA, 5 mM β-mercaptoethanol, 0.1 mM phenylmethylsulfonyl fluoride (PMSF)). The homogenate was crosslinked with 1% formaldehyde at room temperature, and then 0.125 M glycine was added to stop the crosslinking. After filtering the lysate, it was loaded onto dense sucrose buffer (1.7 M sucrose, 20 mM Tris-HCl, 2 mM MgCl₂, 0.2% Triton X-100, 2 mM EDTA, 15 mM β-mercaptoethanol, 0.1 mM PMSF), centrifuged at 2,200g for 20 min at 4 °C, washed twice with 1× NEB2 buffer and resuspended in 1× DpnII buffer. The nuclei were treated with 1% SDS at 65 °C for 10 min, immediately followed by the addition of 10% Triton X-100 to stop the treatment, and then the chromatin was digested overnight with DpnII at 37 °C. After heat inactivation at 65 °C, T4 ligase was added for proximal ligation at 16 °C for 8 h. The crosslinks were reversed overnight with proteinase K at 65 °C, and the DNA was purified using the DNA Clean & Concentrator-5 kit (D4014, Zymo Research). DNA fragments larger than 1 kb were selected and sequenced using the Nanopore Ligation Sequencing kit (SQK-LSK109, ONT). Hi-C libraries were prepared following the method described previously⁵¹ with minor modifications. In brief, calluses were vacuum infiltrated in nuclei isolation buffer supplemented with 2% formaldehyde for 15 min, ground with liquid nitrogen and washed sequentially with extraction buffer I (0.4 M sucrose, 10 mM Tris-HCl pH 8.0, 10 mM MgCl₂, 5 mM β-mercaptoethanol, 0.1 mM PMSF and protease inhibitor), extraction buffer II (0.25 M sucrose, 10 mM Tris-HCl, 10 mM MgCl₂, 1% Triton X-100, 5 mM β-mercaptoethanol, 0.1 mM PMSF and protease inhibitor) and extraction buffer III (1.7 M sucrose, 10 mM Tris-HCl, 0.15% Triton X-100, 2 mM MgCl₂, 5 mM β-mercaptoethanol, 0.1 mM PMSF and

protease inhibitor). Overnight digestion was performed with restriction enzyme (400 units DpnII) at 37 °C on a rocking platform. The constructed Hi-C libraries were sequenced on a DNBSEQ-T7 platform (BGI).

A total of 215 *Rosa* accessions representing wild species, intermediate old cultivars and modern cultivars were selected for genome resequencing. High-quality genomic DNA was extracted from young fresh leaves using the CTAB⁴⁹ method. Illumina paired-end libraries with insert sizes of ~250 bp were constructed using the VAHTS Universal Plus DNA library prep kit for Illumina (Vazyme) following manufacturer recommendations and sequenced on an Illumina NovaSeq 6000 platform (Illumina).

Genome size estimation and ploidy analysis

The genome size of 'Samantha' was estimated using flow cytometry⁵². Leaves were placed in a 500 µl CyStain PI Absolute P nuclei extraction buffer (Sysmex Partec), chopped with a razor blade and then filtered through a 50-µm filter. The collected cells were combined with 2,000 µl of CyStain PI Absolute P staining buffer (Sysmex Partec) and incubated in the dark for 30 min. The nuclei suspension was analysed using a CyFlow Space flow cytometer (Sysmex Partec). Genome size estimations were independently conducted using *Solanum lycopersicum* cv. Heinz 1706 as the internal reference standard.

Ploidy analysis was conducted with young fresh leaf tissue, using the same cell extraction approach described above for the genome size estimation. CyStain UV Precise P staining buffer was added and samples were incubated at room temperature for 1 min. The ploidy level of each sample was analysed using a CyFlow ploidy analyser (Sysmex Partec). FACS Express (v.3) was used for gating (Supplementary Fig. 8) and data analysis (Supplementary Fig. 9).

RNA extraction and transcriptome sequencing

Total RNA was extracted from floral buds, fruits, leaves, petals, pistils, stamens, roots, seeds, stems and thorns of 'Samantha' using the hot borate method³⁴ with modifications. Briefly, the tissues were ground in liquid nitrogen and homogenized with a preheated extraction buffer (200 mM sodium tetraborate decahydrate, 30 mM EGTA, 1% deoxy-cholic acid sodium salt, 10 mM dithiothreitol, 2% PVP 40, 1% NP-40) at 2.5 mg g⁻¹ fresh weight. Subsequently, 25 µl of 25 mg ml⁻¹ protease K (Sigma-Aldrich) was added and the mixture incubated at 42 °C for 2 h. Afterwards, 0.1 ml of 2 M KCl was introduced and the mixture incubated at 4 °C for 1 h. The sample was then centrifuged at 4 °C for 30 min and the supernatant transferred to a 2 ml RNase-free centrifuge tube. One-third volume of 8 M LiCl was then added to the supernatant and the mixture incubated overnight at 4 °C. The following day, the sample was centrifuged at 12,000g for 30 min at 4 °C. The resulting pellet was washed with 1 ml of 2 M LiCl, dissolved in 380 µl of Tris-HCl (pH 7.5) and added with 38 µl potassium acetate. The mixture was then incubated at 4 °C for 15 min, and the supernatant was transferred to a fresh 1.5 ml RNase-free centrifuge tube, followed by the addition of 1 ml ethanol. The mixture was incubated at -80 °C for 2 h to precipitate the RNA. Finally, the pellet was washed with 70% ethanol and dissolved in RNase-free water, yielding purified RNA. Strand-specific RNA-seq libraries were constructed using the VAHTS Universal V10 RNA-seq library prep kit (Vazyme) and sequenced on the DNBSEQ-T7 platform (BGI). Three biological replicates were conducted for each sample.

De novo genome assembly of 'Samantha'

ONT ultra-long reads, PacBio HiFi reads and Hi-C reads were de novo assembled into contigs using hifiasm⁵³ (v.0.19.8-r603) with parameters ' $D = 10$, $r = 4$ and $a = 5$ '. Potential contaminations from microorganisms were then detected in the assembly by aligning the assembled contigs to the GenBank common eukaryotic contaminant database (https://ftp.ncbi.nlm.nih.gov/pub/kitts/contam_in_euks.fa.gz) and the GenBank nt database (<https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nt.gz>) using BLASTN⁵⁴ (v.2.12.0+) with an *e*-value cut-off of 1×10^{-5} . Contigs with more

than 90% of sequence identity to sequences from microorganisms were considered contaminants and removed. Pore-C reads were then utilized to assemble the contigs into 28 pseudochromosomes. First, the DpnII restriction sites on the contigs were analysed using Cooler⁵⁵ (v.0.8.6). Subsequently, Pore-C reads were aligned to the contigs using minimap2 (v.2.17-r941)⁵⁶, and fragment annotation and generation of the pairwise contact matrix were carried out using the HiPore-C pipeline (<https://github.com/zhengdafangyuan/HIPore-C>). The pairwise contact matrix was used to anchor contigs using LACHESIS⁵⁷. To identify potential collapsed regions in the assembled chromosomes, PacBio HiFi reads were aligned to the assembled genome using minimap2 (v.2.17-r941)⁵⁶ with default parameters. On the basis of the alignments, the mapping depth was calculated for each 100-kb non-overlapping window across the assembly using mosdepth (v.0.3.6)⁵⁸. Genome regions with a depth greater than 100 \times (twice the normal depth of 50 \times) were identified as potential collapsed regions. The assembled contigs were then aligned to these regions using minimap2 (v.2.17-r941)⁵⁶ to obtain the corresponding contigs, which were then re-anchored to the chromosomes using HiCUP (v.0.6.1)⁵⁹. The final haplotype-resolved chromosome assembly results were manually corrected using JuiceBox (v.1.11.08)⁶⁰ on the basis of Hi-C read mapping, resulting in a total of 28 haplotype-resolved chromosomes. The accuracy of phasing in the 'Samantha' genome assembly was assessed using nPhase (v.1.2.0)¹³. Synteny between the assembly and the *R. chinensis* 'Old Blush' genome⁸ was identified using SYRI (v.1.7.0)⁶¹, and synteny between the assembly and the K5 genetic map of modern rose¹² was identified using ALLMAPS (v.0.8.12)⁶². BUSCO (v.5.4.5)¹⁰ and LAI¹¹ (LTR_retriever; v.2.9.0) were used to evaluate the quality of the assembly. *K*-mer spectrum analysis was conducted using the KAT programme (v.2.4.2)⁶³. Merqury (v.1.3)⁶⁴ was further utilized to assess the consensus quality value and *k*-mer completeness of the assembly.

Repeat annotation and protein-coding gene prediction

A de novo long terminal repeat retrotransposon (LTR-RT) library and a miniature inverted repeat transposable element (MITE) library were constructed for the assembled 'Samantha' genome using LTRharvest (v.1.5.10)⁶⁵ and MITE-Hunter (v.11-2011)⁶⁶, respectively. The 'Samantha' assembly was then masked with the LTR-RT and MITE libraries using RepeatMasker (v.4.0.6)⁶⁷. The unmasked sequences in the genome were further searched for repeat elements using RepeatModeler (v.1.0.11)⁶⁸. All the repetitive sequences generated above were combined into one de novo repeat library, and along with the Repbase (v.29.04)⁶⁹ and Dfam (v.3.8)⁷⁰ databases, were used to search the genome for repeat sequences using RepeatMasker (v.4.0.6)⁶⁷.

The repeat-masked genome was used for protein-coding gene prediction with the EVidenceModeler pipeline (v.2.1.0)⁷¹, which combines evidence of ab initio gene prediction, transcript mapping and protein homology to define the final gene models. SNAP (v.2006-07-28)⁷², GlimmerHMM (v.3.0.4)⁷³ and AUGUSTUS (v.3.3)⁷⁴ were used for ab initio gene predictions. The paired-end RNA-seq data were cleaned with Trimmomatic (v.0.39)⁷⁵ and assembled using Trinity (v.2.4.0)⁷⁶ and PASA (v.2.5.3)⁷⁷, with the de novo mode and the genome-guided mode, respectively. The assembled transcripts were used as the transcript evidence. Protein sequences from *Rubus*⁷⁸, *Malus*⁷⁹, *Fragaria*⁸⁰ and diploid rose^{8,16-18} were aligned to the 'Samantha' genome using miniprot (v.0.13)⁸¹ to provide the protein homology evidence. For gene function annotation, protein sequences of the predicted genes were compared against the *Arabidopsis* protein⁸² and UniProt⁸³ (Swiss-Prot/TrEMBL) (v.2023_01) databases using BLASTP (v.2.12.0+)⁵⁴, as well as the InterPro database using InterProScan (v.5.0)⁸⁴. GO annotations were obtained using Blast2GO (v.6.0)⁸⁵. NLR-Annotator (v.2.1b)⁸⁶ was utilized for the identification of NLR (nucleotide-binding leucine-rich repeat) receptors.

Variant calling and annotation

The Illumina paired-end reads from each rose accession were processed to remove adaptor and low-quality sequences using Trimmomatic

(v.0.39)⁷⁵. ChrA and ChrB were selected to represent the ‘Samantha’ genome for variant calling. The cleaned reads were aligned to ChrA and ChrB using BWA-MEM (v.0.7.17)⁸⁷ with default parameters. Picard (v2.7.1; <http://broadinstitute.github.io/picard/>) was used to mark duplicated alignments. The HaplotypeCaller function in GATK (v.4.2.6)⁸⁸ was then used to generate a VCF file for each of the 233 rose accessions. All VCF files were combined for variant calling using the function GenotypeGVCFs in GATK⁸⁸. Hard filter was applied to the identified raw SNPs using GATK with the parameters ‘QD < 2.0 || FS > 60.0 || MQ < 40.0’. SNPs were further filtered to retain only those with biallelic sites, a minor allele frequency (MAF) of >0.05 and a data missing rate of <10%. For subsequent SNP analysis, the scoring was based on codominant markers. SNP annotation was performed using the package ANNOVAR (v.2015-12-14)⁸⁹.

Phylogenetic and population genomic analysis

A total of 501,064 SNPs at the 4DTV (4-fold synonymous third-codon transversion) sites were used for phylogenetic and population structure analyses. A maximum-likelihood phylogenetic tree was constructed using FastTree (v.2.1.11)⁹⁰ with 1,000 bootstraps. PCA was performed using PLINK (v.1.90b6.10)⁹¹. Population structure was investigated using ADMIXTURE (v.1.3.0)⁹² with 200 bootstraps. To determine the most likely number of ancestral kinships (K) in the rose population, K values were set from 2 to 10. The statistic ‘cross-validation error’ (CV error), which indicates the change in likelihood of different numbers of clusters, was calculated, and the cluster number with the lowest CVerror indicating the most likely number of clusters in the population was obtained.

Nucleotide diversity (π) was calculated in 500-kb non-overlapping windows across the ‘Samantha’ genome with the final set of 17,760,579 SNPs using VCFtools (v.0.1.16)⁹³. Fixation index (F_{ST}) values were calculated between two groups (tetraploid intermediate old cultivars and tetraploid modern cultivars) using VCFtools (v.0.1.16)⁹³.

Linkage disequilibrium decay patterns were calculated using PopLDdecay (v.3.42)⁹⁴ with default parameters for the two groups (tetraploid intermediate old cultivars and tetraploid modern cultivars). The correlation coefficient (r^2) of genotypes was used as a measure of the linkage disequilibrium level. SNPs within each group were extracted for the analysis. SNPs within 2-kb sliding windows were used to estimate the average r^2 for various physical distance classes, and the linkage disequilibrium decay was plotted as a function of the derived average r^2 and the physical distances along the genome.

Selective sweep identification

We identified the selection signals across the genome on the basis of the π ratios and F_{ST} values. Selective sweep screening was performed by comparing tetraploid intermediate old cultivars (Int_4; $n = 45$) vs tetraploid modern cultivars (Hyb_4; $n = 32$). A 10-kb sliding window with 1-kb step approach was applied to quantify F_{ST} and π ratios. Regions with the top 5% of both F_{ST} values and π ratios were considered as candidate selective sweeps.

To infer potential sources of interesting selective sweeps in tetraploid modern cultivars, 54 accessions from original species (single genetic component in population structure analysis) were used for SNP genotyping analysis (Supplementary Table 8). SNPs in each selective sweep were used to construct the maximum-likelihood phylogenetic tree with IQ-TREE (v.1.6.8)⁹⁵, with a bootstrap value of 1,000. The genotypes of SNPs in each selective sweep region are presented using a heat map generated with the R package pheatmap (v.1.0.12).

Genomic composition analysis

Six potential original species (*R. chinensis* ‘Old Blush’, *R. odorata* var. *gigantea*, *R. gallica*, *R. wichuraiana*, *R. fedtschenkoana* and *R. moschata*) of the ‘Samantha’ genome identified by ADMIXTURE⁹³ were used for genetic organization analysis. The original attribution of

chromosomes was inferred through two approaches: (1) Equal amounts of Illumina short-read data from each of the six species were aligned to the homologous chromosomes of ‘Samantha’. Within each of the 500-kb non-overlapping windows, the window was considered to be attributed to the species with the highest sequencing coverage depth. (2) High-quality SNPs between each original species and ‘Samantha’ were called on the basis of the alignments, and the potential original attribution of the 500-kb non-overlapping windows was inferred by calculating the variant frequency between each species and the ‘Samantha’ genome using the following formula:

$$\text{Variant frequency} = \frac{\text{number of homozygous variant sites} + \frac{1}{2} \times \text{number of heterozygous sites}}{\text{total number of identified sites}} \quad (1)$$

A lower variant frequency indicates a higher sequence similarity between the potential original species and ‘Samantha’ and a higher probability that the window/genome region originates from the corresponding species. For each specific genomic region of ‘Samantha’, the species with the lowest variant frequency was considered to be the original contributor of this region. Only results obtained by the two approaches that were consistent were used to determine the final contributing species of each region across the ‘Samantha’ chromosomes.

Reporting summary

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

Data availability

Raw reads generated in this study have been deposited in the NCBI BioProject database under accession numbers PRJNA1108167 and PRJNA704782. The sequences and annotations of the ‘Samantha’ genome assembly are available on figshare at <https://doi.org/10.6084/m9.figshare.22774097> (ref. 96). The following databases were used: BUSCO eudicot database (https://busco-data.ezlab.org/v5/data/lineages/eudicots_odb10.2024-01-08.tar.gz), GenBank common eukaryotic contaminant database (https://ftp.ncbi.nlm.nih.gov/pub/kitts/contam_in_euks.fa.gz), GenBank nt database (<https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nt.gz>), InterPro database (<https://www.ebi.ac.uk/interpro/download/InterPro/>), UniProt (Swiss-Prot/TrEMBL) databases (<https://www.uniprot.org/downloads>), Repbase database (<https://www.girinst.org/downloads/>), Dfam database (https://www.dfam.org/releases/Dfam_3.8/families/), *Fragaria ananassa* ‘Yanli’ genome (<https://www.rosaceae.org/Analysis/14723107>), *Malus domestica* ‘Fuji’ genome (<https://www.rosaceae.org/Analysis/15540493>), *Rubus idaeus* ‘Joan’ genome (<https://www.rosaceae.org/Analysis/14031373>), *Rosa rugosa* genome (<https://www.rosaceae.org/Analysis/11775539>), *Rosa chinensis* ‘Old Blush’ genome (<https://www.rosaceae.org/Analysis/282>), *Rosa wichuraiana* ‘Basye’s Thornless’ genome (<https://www.rosaceae.org/Analysis/13087667>) and *Rosa chinensis* ‘Chilong Han-zhu’ genome (<https://www.ncbi.nlm.nih.gov/bioproject/932466>; <https://doi.org/10.6084/m9.figshare.26888665.v1>).

References

- Usman, M., Ashfaq, M., Taj, S. & Abid, M. An economic analysis of cut-rose flower in Punjab, Pakistan. *J. Anim. Plant Sci.* **24**, 651–655 (2014).
- Roberts, A. V., Debener, T. & Gudin, S. *Encyclopedia of Rose Science* (Academic Press, 2003).
- Yokoya, K., Roberts, A., Mottley, J., Lewis, R. & Brandham, P. Nuclear DNA amounts in roses. *Ann. Bot.* **85**, 557–561 (2000).
- Wylie, A. P. The history of garden roses. *J. R. Hortic. Soc.* **79**, 555–571 (1954).

5. de Vries, D. & Dubois, L. Rose breeding: past, present, prospects. *Acta Hortic.* **424**, 241–248 (1996).
6. Bendahmane, M., Dubois, A., Raymond, O. & Bris, M. L. Genetics and genomics of flower initiation and development in roses. *J. Exp. Bot.* **64**, 847–857 (2013).
7. Hibrand Saint-Oyant, L. et al. A high-quality genome sequence of *Rosa chinensis* to elucidate ornamental traits. *Nat. Plants* **4**, 473–484 (2018).
8. Raymond, O. et al. The *Rosa* genome provides new insights into the domestication of modern roses. *Nat. Genet.* **50**, 772–777 (2018).
9. Deshpande, A. S. et al. Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing. *Nat. Biotechnol.* **40**, 1488–1499 (2022).
10. Simão, 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).
11. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
12. Bourke, P. M. et al. Partial preferential chromosome pairing is genotype dependent in tetraploid rose. *Plant J.* **90**, 330–343 (2017).
13. Abou Saada, O., Tsouris, A., Eberlein, C., Friedrich, A. & Schacherer, J. nPhase: an accurate and contiguous phasing method for polyploids. *Genome Biol.* **22**, 126 (2021).
14. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
15. 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).
16. 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).
17. 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).
18. Zhong, M.-C. et al. Rose without prickle: genomic insights linked to moisture adaptation. *Natl Sci. Rev.* **8**, nwab092 (2021).
19. Soufflet-Freslon, V. et al. Diversity and selection of the continuous-flowering gene, *RokSN*, in rose. *Hortic. Res.* **8**, 76 (2021).
20. Iwata, H. et al. The *TFL1* homologue *KSN* is a regulator of continuous flowering in rose and strawberry. *Plant J.* **69**, 116–125 (2011).
21. Luo, X. & He, Y. Experiencing winter for spring flowering: a molecular epigenetic perspective on vernalization. *J. Integr. Plant Biol.* **62**, 104–117 (2020).
22. Weingartner, M., Subert, C. & Sauer, N. LATE, a C₂H₂ zinc-finger protein that acts as floral repressor. *Plant J.* **68**, 681–692 (2011).
23. Sharma, N. et al. NO FLOWERING IN SHORT DAY (NFL) is a bHLH transcription factor that promotes flowering specifically under short-day conditions in *Arabidopsis*. *Development* **143**, 682–690 (2016).
24. Xu, M. et al. Developmental functions of miR156-regulated SQUAMOSA PROMOTER BINDING PROTEIN-LIKE (SPL) genes in *Arabidopsis thaliana*. *PLoS Genet.* **12**, e1006263 (2016).
25. Clark, S. E., Running, M. P. & Meyerowitz, E. M. CLAVATA1, a regulator of meristem and flower development in *Arabidopsis*. *Development* **119**, 397–418 (1993).
26. DeYoung, B. J. et al. The CLAVATA1-related BAM1, BAM2 and BAM3 receptor kinase-like proteins are required for meristem function in *Arabidopsis*. *Plant J.* **45**, 1–16 (2005).
27. Saddic, L. A. The LEAFY target LMI1 is a meristem identity regulator and acts together with LEAFY to regulate expression of CAULIFLOWER. *Development* **133**, 1673–1682 (2006).
28. François, L. et al. A miR172 target-deficient AP2-like gene correlates with the double flower phenotype in roses. *Sci. Rep.* **8**, 12912 (2018).
29. 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).
30. Han, Y. et al. An APETALA2 homolog, RcAP2, regulates the number of rose petals derived from stamens and response to temperature fluctuations. *Front. Plant Sci.* **9**, 481 (2018).
31. Grotewold, E. The genetics and biochemistry of floral pigments. *Annu. Rev. Plant Biol.* **57**, 761–780 (2006).
32. Milbrath, J., Hansen, E. & Hartman, H. Defoliation of rose plants with ethylene gas. *Science* **91**, 100 (1940).
33. Reid, M. S., Evans, R. Y., Dodge, L. L. & Mor, Y. Ethylene and silver thiosulfate influence opening of cut rose flowers. *J. Am. Soc. Hortic. Sci.* **114**, 436–440 (1989).
34. Ma, N. et al. Transcriptional regulation of ethylene receptor and CTR genes involved in ethylene-induced flower opening in cut rose (*Rosa hybrida*) cv. Samantha. *J. Exp. Bot.* **57**, 2763–2773 (2006).
35. Alonso, J. M. & Stepanova, A. N. The ethylene signaling pathway. *Science* **306**, 1513–1515 (2004).
36. Gong, Z. et al. ROS1, a repressor of transcriptional gene silencing in *Arabidopsis*, encodes a DNA glycosylase/lyase. *Cell* **111**, 803–814 (2002).
37. Harada, T. et al. Cloning, characterization, and expression of xyloglucan endotransglucosylase/hydrolase and expansin genes associated with petal growth and development during carnation flower opening. *J. Exp. Bot.* **62**, 815–823 (2011).
38. Cai, X. et al. Mutant identification and characterization of the laccase gene family in *Arabidopsis*. *J. Exp. Bot.* **57**, 2563–2569 (2006).
39. Zhao, Q. et al. LACCASE is necessary and nonredundant with PEROXIDASE for lignin polymerization during vascular development in *Arabidopsis*. *Plant Cell* **25**, 3976–3987 (2013).
40. Zhang, J. et al. Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm. *BMC Genomics* **16**, 217 (2015).
41. Cao, K. et al. Genome-wide association study of 12 agronomic traits in peach. *Nat. Commun.* **7**, 13246 (2016).
42. Debener, T. & Byrne, D. H. Disease resistance breeding in rose: current status and potential of biotechnological tools. *Plant Sci.* **228**, 107–117 (2014).
43. Hackenberg, T. et al. Catalase and NO CATALASE ACTIVITY1 promote autophagy-dependent cell death in *Arabidopsis*. *Plant Cell* **25**, 4616–4626 (2013).
44. Hu, Y. et al. Ectopic expression of *Arabidopsis* broad-spectrum resistance gene RPW8.2 improves the resistance to powdery mildew in grapevine (*Vitis vinifera*). *Plant Sci.* **267**, 20–31 (2018).
45. Fang, P. et al. Analysis of allelic variants of RhMLO genes in rose and functional studies on susceptibility to powdery mildew related to clade V homologs. *Theor. Appl. Genet.* **134**, 2495–2515 (2021).
46. Caplan, J., Padmanabhan, M. & Dinesh-Kumar, S. P. Plant NB-LRR immune receptors: from recognition to transcriptional reprogramming. *Cell Host Microbe* **3**, 126–135 (2008).
47. Satterlee, J. W. et al. Convergent evolution of plant prickles by repeated gene co-option over deep time. *Science* **385**, eado1663 (2024).
48. Scalliet, G. et al. Scent evolution in Chinese roses. *Proc. Natl Acad. Sci. USA* **105**, 5927–5932 (2008).

49. Allen, G. C., Flores-Vergara, M. A., Krasynanski, S., Kumar, S. & Thompson, W. F. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nat. Protoc.* **1**, 2320–2325 (2006).
50. Li, Z. et al. Pore-C simultaneously captures genome-wide multi-way chromatin interaction and associated DNA methylation status in *Arabidopsis*. *Plant Biotechnol. J.* **20**, 1009–1011 (2022).
51. Belton, J.-M. et al. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).
52. Hare, E. E. & Johnston, J. S. Genome size determination using flow cytometry of propidium iodide-stained nuclei. *Methods Mol. Biol.* **772**, 3–12 (2011).
53. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
54. Boratyn, G. M. et al. Domain enhanced lookup time accelerated BLAST. *Biol. Direct* **7**, 12 (2012).
55. Abdennur, N. & Mirny, L. A. Cooler: scalable storage for Hi-C data and other genetically labeled arrays. *Bioinformatics* **36**, 311–316 (2019).
56. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
57. Burton, J. N. et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
58. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
59. Wingett, S. W. et al. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research* **4**, 1310 (2015).
60. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
61. Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
62. Tang, H. et al. ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol.* **16**, 3 (2015).
63. Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J. & Clavijo, B. J. KAT: a K mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* **33**, 574–576 (2017).
64. 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).
65. Ellinghaus, D., Kurtz, S. & Willhoefft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
66. Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* **38**, e199 (2010).
67. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* <https://doi.org/10.1002/0471250953.bi0410s25> (2009).
68. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci. USA* **117**, 9451–9457 (2020).
69. Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
70. Hubley, R. et al. The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**, D81–D89 (2015).
71. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
72. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
73. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
74. Stanke, M., Tzvetkova, A. & Morgenstern, B. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol.* **7**, S11 (2006).
75. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
76. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
77. Haas, B. J. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
78. Zhou, J. et al. Comparison of red raspberry and wild strawberry fruits reveals mechanisms of fruit type specification. *Plant Physiol.* **193**, 1016–1035 (2023).
79. Daccord, N. et al. High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nat. Genet.* **49**, 1099–1106 (2017).
80. Mao, J. et al. High-quality haplotype-resolved genome assembly of cultivated octoploid strawberry. *Hortic. Res.* **10**, uhad002 (2023).
81. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, btad014 (2023).
82. Cheng, C.-Y. et al. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J.* **89**, 789–804 (2017).
83. Bairoch, A. The universal protein resource (UniProt). *Nucleic Acids Res.* **33**, D154–D159 (2005).
84. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
85. Conesa, A. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
86. Steuernagel, B. et al. The NLR-Annotator tool enables annotation of the intracellular immune receptor repertoire. *Plant Physiol.* **183**, 468–482 (2020).
87. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
88. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
89. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
90. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
91. 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).
92. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
93. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
94. 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 (2019).

95. 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).
96. Yang, T. Genome annotation files of *Rosa hybrida* ‘Samantha®’. figshare <https://doi.org/10.6084/m9.figshare.22774097.v1> (2024).

Acknowledgements

We thank T. Lin (China Agricultural University) and Q. Gao (BGI Tech., Beijing) for helpful discussions; H. Xin and Y. Bu (Beijing Institute of Landscape Architecture) for providing computing resources and *Rosa* materials; the Center of Agricultural Biotechnology of Beijing Academy of Agriculture and Forestry Sciences and Golden Intelligence Biotechnology Co., Ltd. for help with flow cytometry experiments; and Wuhan Grandomics Biosciences Co., Ltd. for providing Pore-C sequencing and analysis services. This work was supported by funds from the 111 Project of the Ministry of Education (Grant no. B17043 to J.G.), the Construction of Beijing Science and Technology Innovation and Service Capacity in Top Subjects (Grant no. CEFF-PXM2019_014207_000032 to J.G.), the National Key Research and Development Program of China (Grant no. 2018YFD1000400 to J.G.), the General Project of Shenzhen Science and Technology and Innovation Commission (Grant no. 21K270360620 to Y. Li.), the National Natural Science Foundation of China (Grant no. 32372752 to Y. Li., 31772344 and 31972444 to Z.Z., and 31522049 and 31872148 to N.M.), the National Science Fund for Distinguished Young Scholars of China (Grant no. 32325046 to N.M.), the earmarked fund for CARS (Grant no. CARS-23 to N.M.), and the Postdoctoral Fellowship Program of CPSF (Grant no. GZB20240831 to T.Y.).

Author contributions

J.G., N.M., Z.F. and Z.Z. designed and coordinated the project. Y. Liu., W.W., H.R. and S.S. performed DNA extraction. T.Y. and Y. Liu. performed the flow cytometry analysis. Y.Y., L.L., S.D., Y. Zhu., Y.C., H. Zhou., H. Zhang., J.C. and K.T. contributed *Rosa* materials. Z.Z., Y. Liu., Y. Li., T.Y., Q.P., X.S., Y.J. and X.Z. coordinated sample collection and sequence data generation. T.Y., D.G., L.C., S.W., S.S., H.S., J.W. and Y. Zhang integrated the genome assembly and annotation. T.Y. and

H.S. performed selective sweep analysis. T.Y., H.S., J.W. and Y. Zhang conducted phylogenetic and population genomic analysis. T.Y., Y. Liu., J.W., S.W., Z.Z., Y. Li., Z.F., N.M. and J.G. wrote and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41477-024-01820-x>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41477-024-01820-x>.

Correspondence and requests for materials should be addressed to Zhangjun Fei, Nan Ma or Junping Gao.

Peer review information *Nature Plants* thanks Fabrice Foucher, Paul Arens and Diana Lopez Arias for their contribution to the peer review of this work.

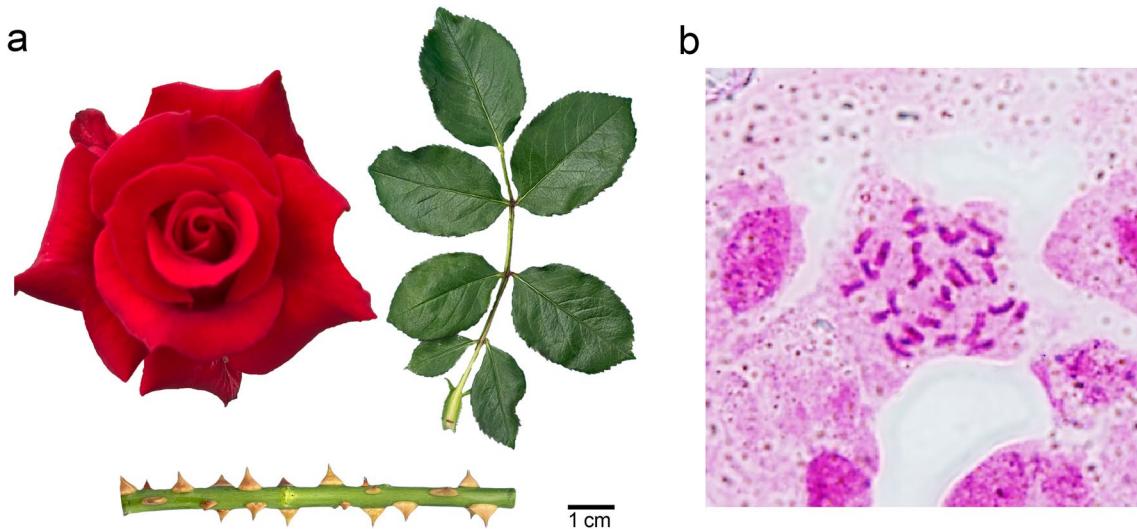
Reprints and permissions information is available at www.nature.com/reprints.

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

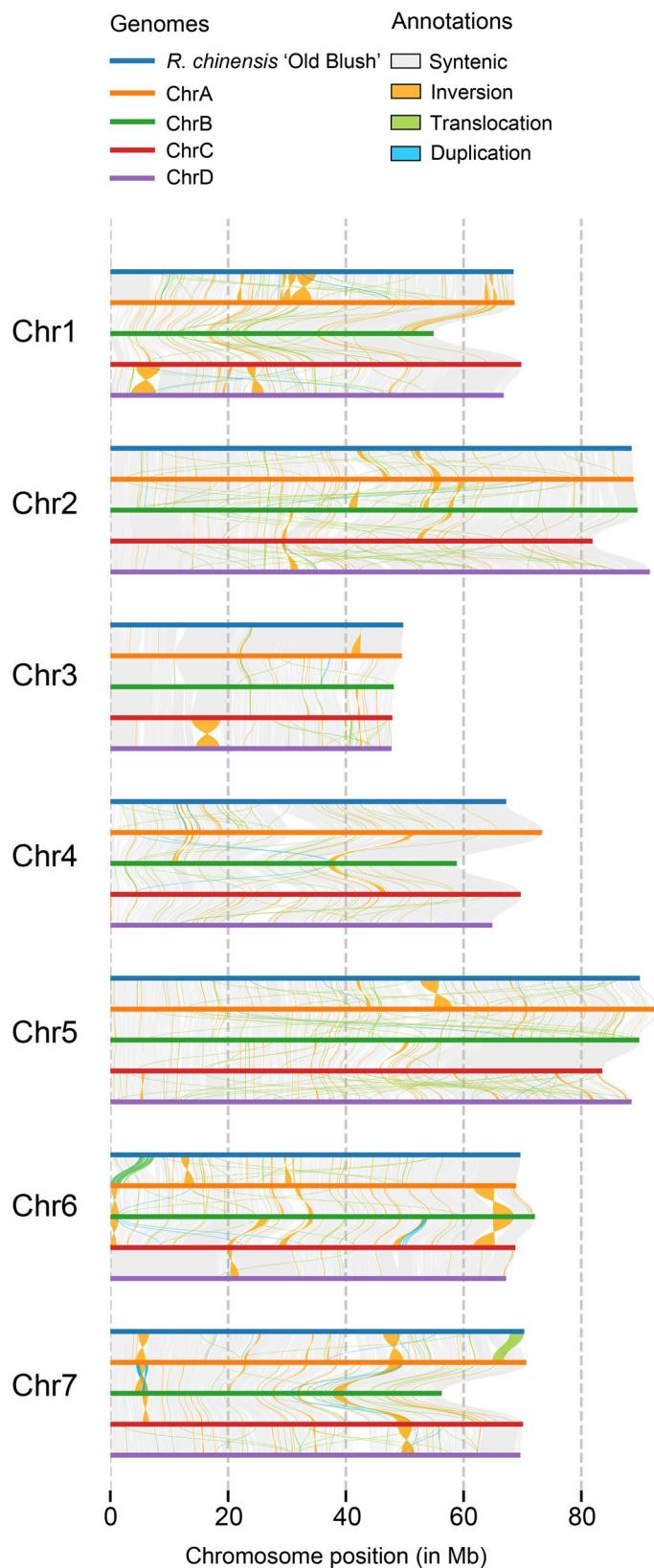
Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing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

© The Author(s), under exclusive licence to Springer Nature Limited 2024

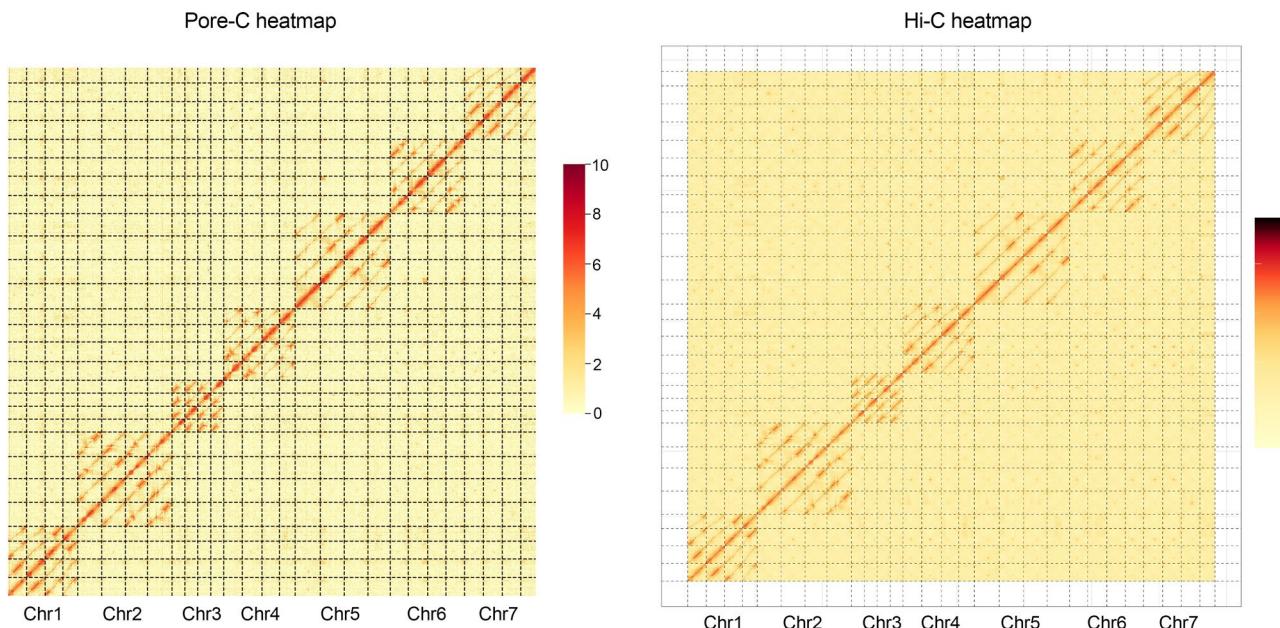
¹Beijing Key Laboratory of Development and Quality Control of Ornamental Crops, Department of Ornamental Horticulture, College of Horticulture, China Agricultural University, Beijing, China. ²Boyce Thompson Institute, Ithaca, NY, USA. ³Plant Biology Section, School of Integrative Plant Science, Cornell University, Ithaca, NY, USA. ⁴School of Food and Drug, Shenzhen Polytechnic University, Shenzhen, Guangdong, China. ⁵Bioinformatics Center, Beijing University of Agriculture, Beijing, China. ⁶College of Plant Science and Technology, Beijing University of Agriculture, Beijing, China. ⁷Kunming Yang Chinese Rose Gardening Co. Ltd., Kunming, Yunnan, China. ⁸People's Government of Weishanzhuang Town, Daxing, Beijing, China. ⁹Chinese Ancient Rose Garden, Beijing, China. ¹⁰Yunnan Xinhaihui Flower Industry Co. Ltd., Tonghai, Yunnan, China. ¹¹Shanghai Academy of Agricultural Sciences, Shanghai, China. ¹²College Horticulture and Landscape Architecture, Zhongkai University of Agriculture and Engineering, Guangzhou, Guangdong, China. ¹³National Engineering Research Center for Ornamental Horticulture, Flower Research Institute of Yunnan Academy of Agricultural Sciences, Kunming, Yunnan, China. ¹⁴Beijing Botanical Garden, Beijing, China. ¹⁵Smartgenomics Technology Institute, Tianjin, China. ¹⁶USDA-ARS Robert W Holley Center for Agriculture and Health, Ithaca, NY, USA. ¹⁷These authors contributed equally: Zhao Zhang, Tuo Yang, Yang Liu, Shan Wu, Honghe Sun, Jie Wu, Yonghong Li. ✉ e-mail: zf25@cornell.edu; ma_nan@cau.edu.cn; gaojp@cau.edu.cn



Extended Data Fig. 1 | Phenotype and chromosome number of *R. hybrida* 'Samantha®'. **a**, Flower, leaf, and stem of 'Samantha®'. **b**, Chromosomes of 'Samantha®' shown in a representative root cell.



Extended Data Fig. 2 | Chromosomal synteny analysis. Genome synteny between *R. hybrida* ‘Samantha[®]’ and *R. chinensis* ‘Old Blush[®]



Extended Data Fig. 3 | Pore-C and Hi-C heatmaps of the assembled *R. hybrida* ‘Samantha®’ genome. The signal intensity of the Pore-C heatmap is expressed as $\log_e(Z+1)$, while the signal intensity of Hi-C heatmap is expressed as $\log_{10}(Z+1)$. Z represents the calculated interaction intensity.

649-*R. multiflora* 'Single Pink'650-*R. multiflora* 'Thornless'

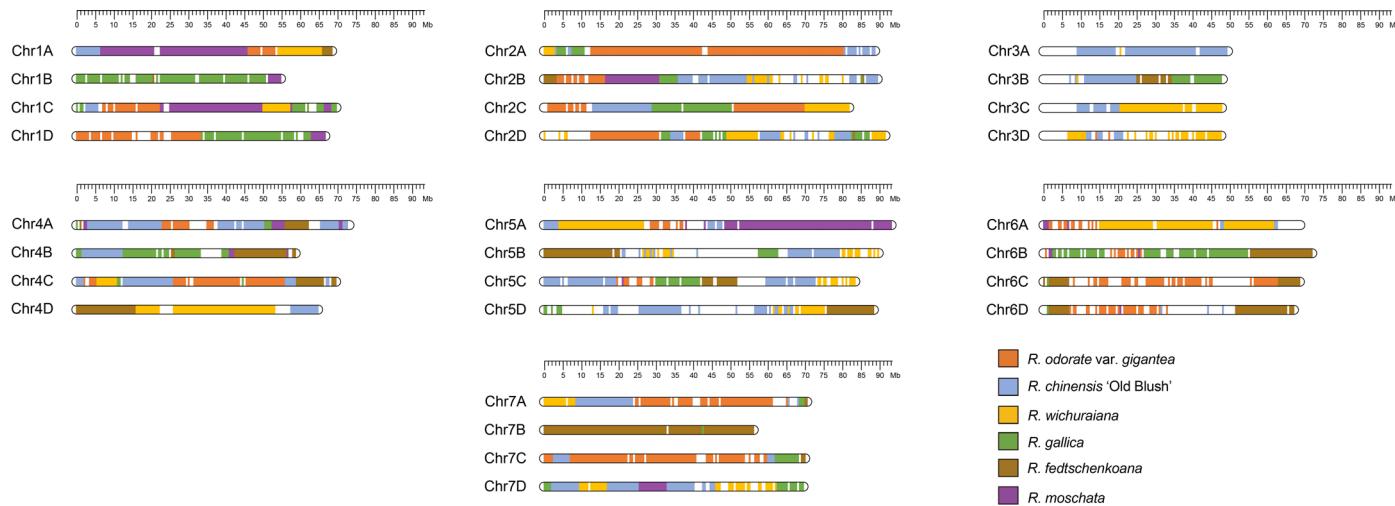
775-Natal Briar (Hybrid Multiflora)



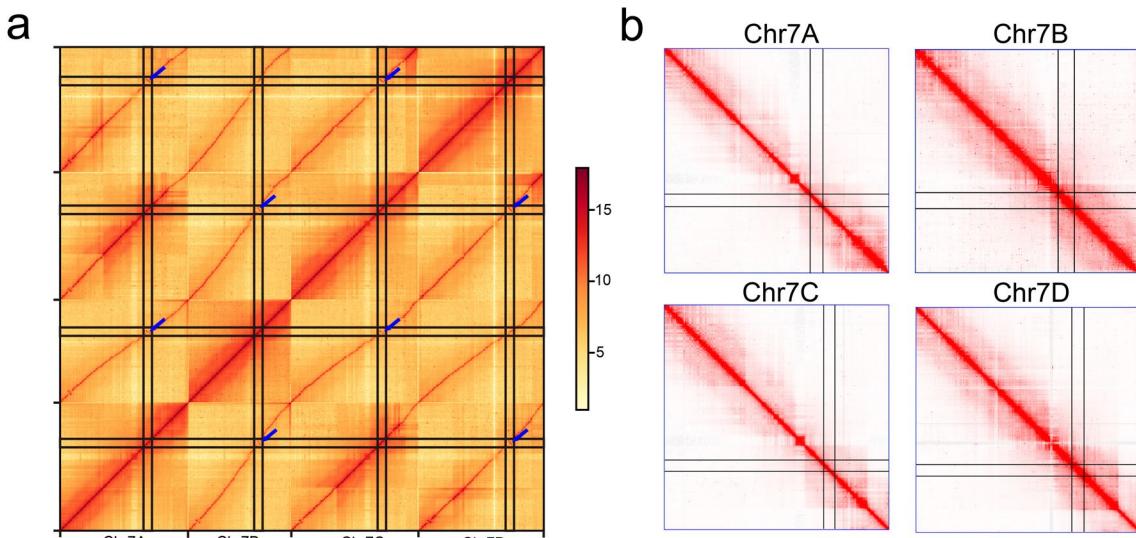
743-Bleu Magenta (Hybrid Multiflora)

775-*R. hybrida* 'Green Ice' (Floribunda)776-*R. hybrida* 'Angela' (Floribunda)

Extended Data Fig. 4 | Inflorescence traits of different rose accessions. The inflorescence traits of the wild species *R. multiflora* have been passed down to Hybrid Multiflora and Floribunda.

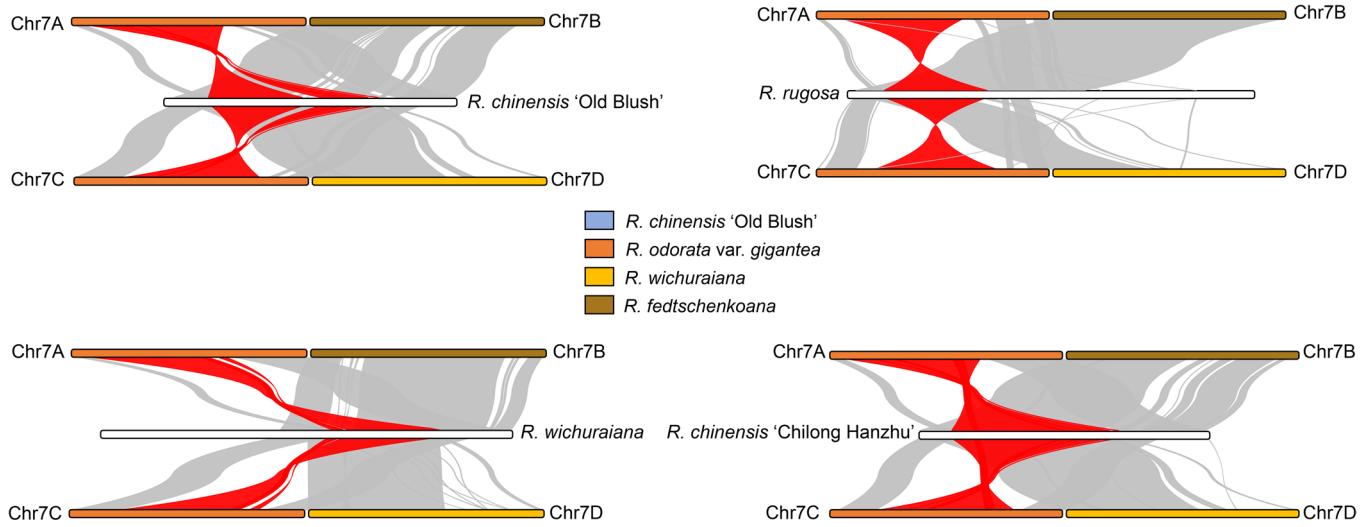


Extended Data Fig. 5 | Genetic organization of the 28 chromosomes of *R. hybrida* 'Samantha®'. Different colors represent genome regions derived from different potential original species.

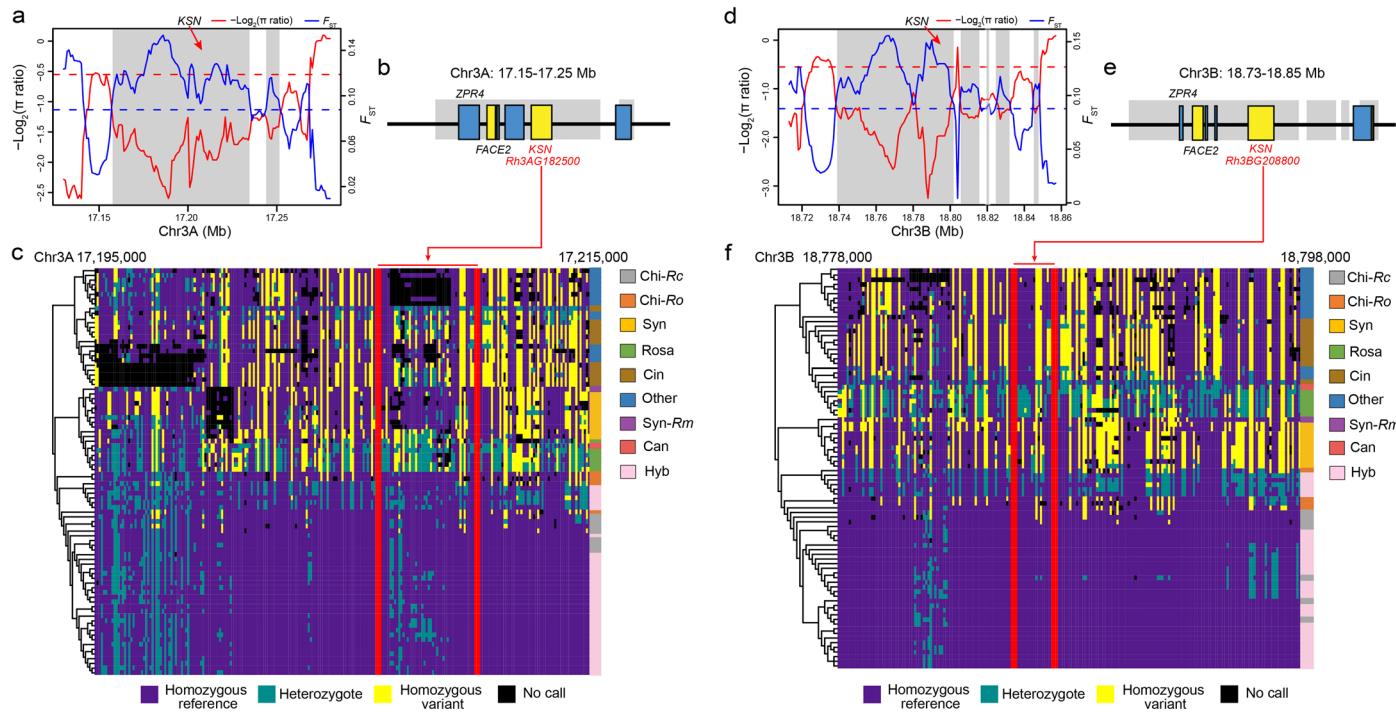


Extended Data Fig. 6 | Pore-C heatmaps showing the inversion sites of the four homologous chromosomes of Chr7. **a**, Pore-C signals between homologous chromosomes of Chr7. Blue arrows indicate the inversion signal between homologous chromosomes. **b**, Pore-C signals in the inversion regions

between homologous chromosomes. Black cross lines indicate the inversion regions. The signal intensity of the Pore-C heatmap is expressed as $\log_e(Z+1)$. Z represents the calculated interaction intensity.

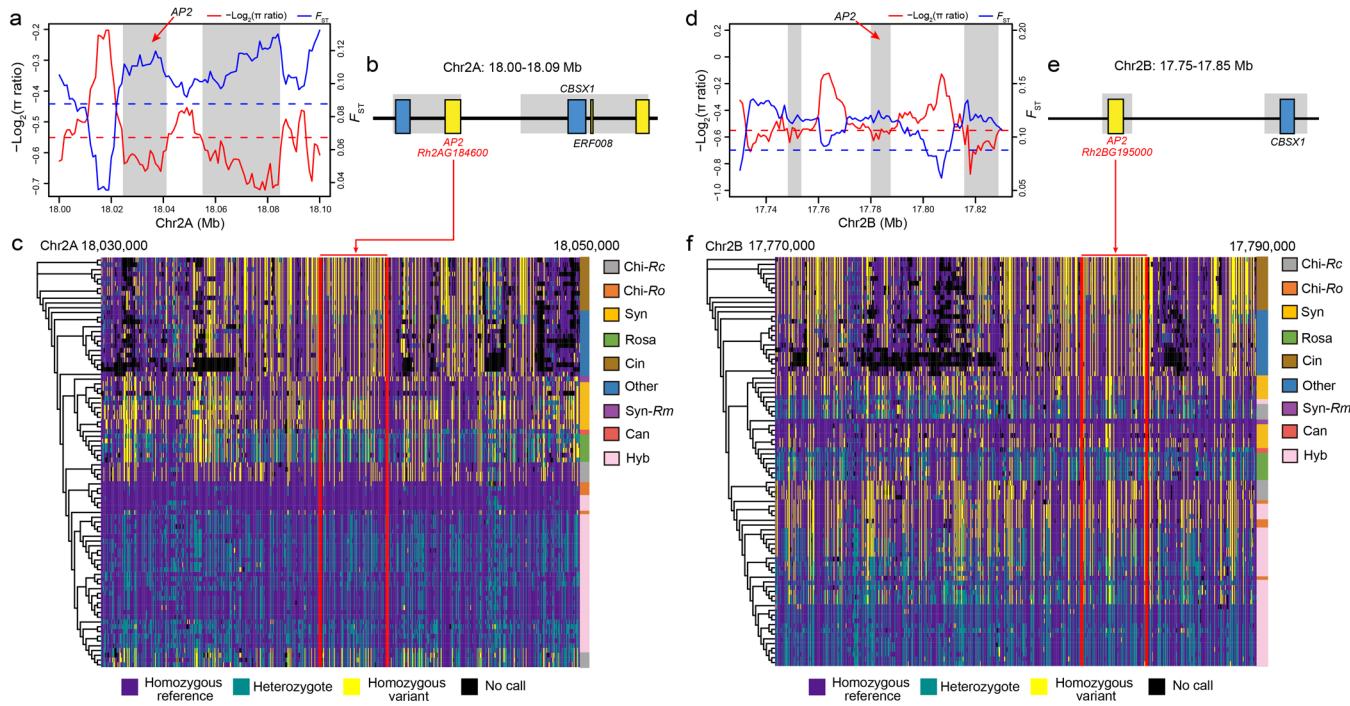


Extended Data Fig. 7 | Collinearity analysis of the region on chromosome 7 harboring an inversion. Four publicly available *Rosa* genomes, *R. chinensis* 'Old Blush'⁸, *R. chinensis* 'Chilong Hanzhu'¹⁶, *R. rugosa*¹⁷ and *R. wichuraiana*¹⁸, were used. The potential origins of this region in the 'Samantha®' genome are indicated with different colors.



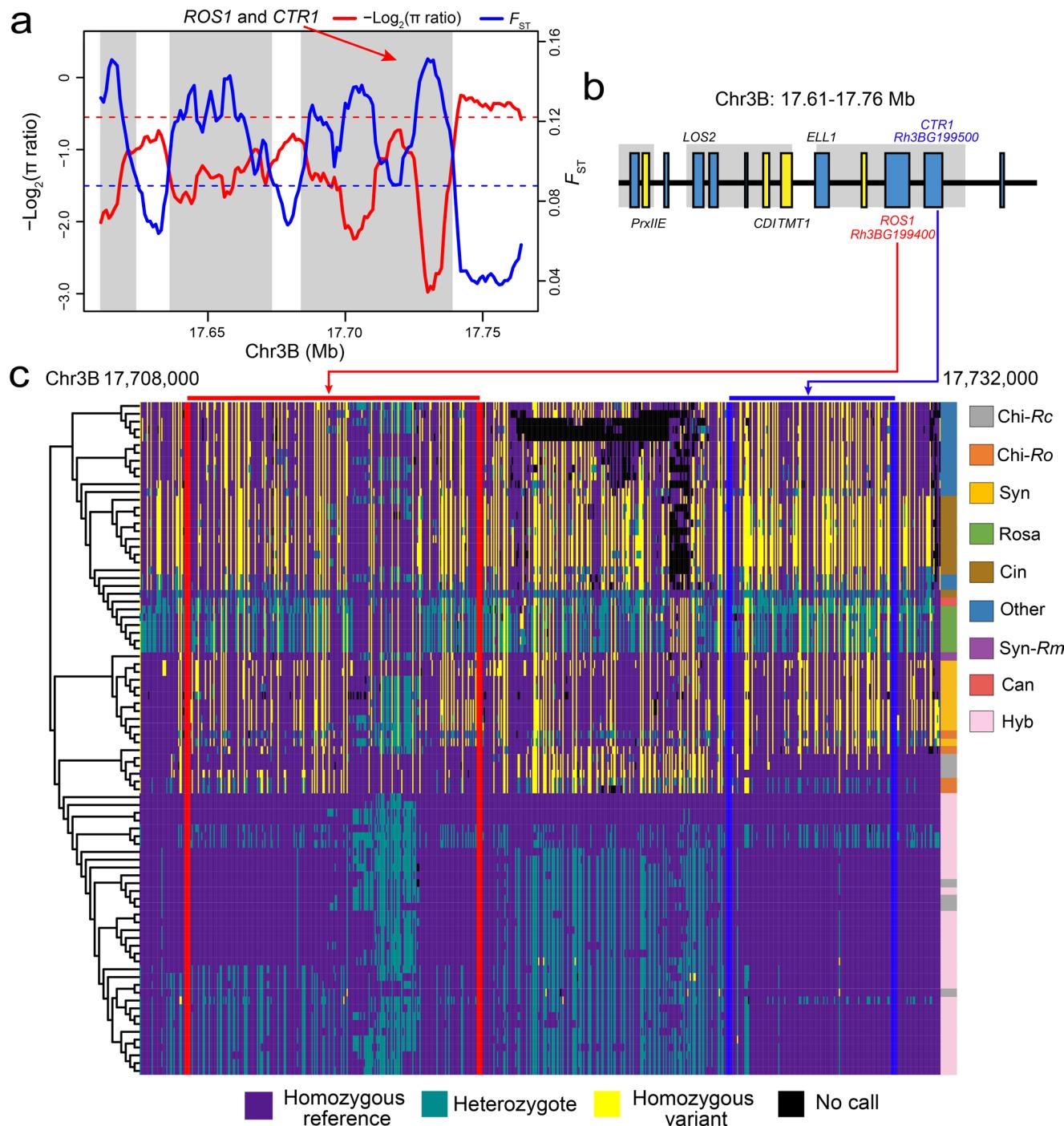
Extended Data Fig. 8 | Selection and evolution of recurrent blooming in modern roses. a,d, π ratio and F_{ST} values in the selective sweep region containing the *KSN* genes. **b,e,** Genes in the selective sweep region. Blue boxes represent genes located on the sense strand, while yellow boxes represent genes located on the antisense strand. **c,f,** Heatmap of SNP genotype profiles in the selective sweep region. Maximum likelihood phylogenetic tree constructed from these

SNPs is shown on the left. The rectangular boxes on the right indicate the sections to which the samples belong. Hyb, modern cultivars, *R. hybrida*; Syn, section *Synstylae*; Syn-Rm, *R. moschata*; Chi-Rc, *R. chinensis* in section *Chinenses*; Chi-Ro, *R. odorata* in section *Chinenses*; Rosa, section *Rosa*; Can, section *Caninae*; Cin, section *Cinnamomeae*; Other, sections *Pimpinellifoliae*, *Microphyllae*, *Bracteatae*, *Banksianae*, and *Laevigatae*.



Extended Data Fig. 9 | Selection and evolution of double flower in modern roses. **a,d**, π ratio and F_{ST} values in the selective sweep region containing the *AP2* genes. **b,e**, Genes in the selective sweep region. Blue boxes represent genes located on the sense strand, while yellow boxes represent genes located on the antisense strand. **c,f**, Heatmap of SNP genotype profiles in the selective sweep region. Maximum likelihood phylogenetic tree constructed from these SNPs

is shown on the left. The rectangular boxes on the right indicate the sections to which the samples belong. Hyb, modern cultivars, *R. hybrida*; Syn, section Synstylae; Syn-Rm, *R. moschata*; Chi-Rc, *R. chinensis* in section Chinenses; Chi-Ro, *R. odorata* in section Chinenses; Rosa, section Rosa; Can, section Caninae; Cin, section Cinnamomeae; Other, sections Pimpinellifoliae, Microphyllae, Bracteatae, Banksianae, and Laevigatae.



Extended Data Fig. 10 | Selection and evolution of ethylene sensitivity in modern roses. **a**, π ratio and F_{ST} values in the selective sweep region containing *CTR1* and *ROS1* genes. **b**, Genes in the selective sweep region. Blue boxes represent genes located on the sense strand, while yellow boxes represent genes located on the antisense strand. **c**, Heatmap of SNP genotype profiles in selective sweep regions. Maximum likelihood phylogenetic tree is shown on the left.

The rectangular boxes on the right indicate the sections to which the samples belong. Hyb, modern cultivars, *R. hybrida*; Syn, section *Synstylae*; Syn-Rm, *R. moschata*; Chi-Rc, *R. chinensis* in section Chinenses; Chi-Ro, *R. odorata* in section Chinenses; Rosa, section Rosa; Can, section Caninae; Cin, section Cinnamomeae; Other, sections Pimpinellifoliae, Microphyllae, Bracteatae, Banksianae, and Laevigatae.

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 commercial and custom code was used for data collection.
Data analysis	Freely available bioinformatics programs were used, including: hifiasm (v.0.19.8-r603), BLASTN (v.2.12.0+), Cooler (v.0.8.6), minimap2 (v.2.17-r941), LACHESIS, JuiceBox (v.1.11.08), HiCUP (v.0.6.1), BUSCO (v.5.4.5), ALLMAPS (v.0.8.12), LTR_retriever (v.2.9.0), nPhase (v.1.2.0), KAT (v.2.4.2), SYRI (v.1.7.0), Merqury (v.1.3), LTRharvest (v.1.5.10), MITE-Hunter (v.11-2011), RepeatMasker (v.4.0.6), RepeatModeler (v.1.0.11), EVidenceModeler (v.2.1.0), SNAP (v.2006-07-28), Repbase (v.29.04), Dfam (v.3.8), AUGUSTUS (v.3.3), GlimmerHMM (v.3.0.4), Trimmomatic (v.0.39), Trinity (v.2.4.0), InterProScan (v.5.0), Blast2GO (v.6.0), UniProt (v.2023_01), NLR-Annotator (v.2.1b), miniprot (v.0.13), PASA (v2.5.3), BWA-MEM (v.0.7.17), Picard (v.2.7.1), GATK (v.4.2.6), ANNOVAR (v.2015-12-14), VCFtools (v.0.1.16), FastTree (v. 2.1.11), PLINK (v.1.90b6.10), ADMIXTURE (v.1.3.0), PopLDdecay (v.3.42), IQ-TREE (v.1.6.8), pheatmap (v.1.0.12), HiPore-C pipeline, FACS Express (v.3), mosdepth (v.0.3.6), BLASTP (v.2.12.0+).

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](#)

Raw reads generated in this study have been deposited in NCBI BioProject database under accession numbers PRJNA1108167 (<https://www.ncbi.nlm.nih.gov/sra?term=PRJNA1108167>) and PRJNA704782 (<https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA704782>). The sequences and annotations of the 'Samantha®' genome assembly are available at Figshare (<https://doi.org/10.6084/m9.figshare.22774097>). BUSCO eudicot database (https://busco-data.ezlab.org/v5/data/lineages/eudicots_odb10.2024-01-08.tar.gz), GenBank common eukaryotic contaminants database (https://ftp.ncbi.nlm.nih.gov/pub/kitts/contam_in_euks.fa.gz), GenBank nt database (<https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nt.gz>), InterPro database (<https://www.ebi.ac.uk/interpro/download/InterPro/>), UniProt (Swiss-Prot/TrEMBL) databases (<https://www.uniprot.org/downloads>), Repbase database (<https://www.girinst.org/downloads/>), Dfam database (https://www.dfam.org/releases/Dfam_3.8/families/), Fragaria ananassa 'Yanli' genome (<https://www.rosaceae.org/Analysis/14723107>); Malus domestica 'Fuji' genome (<https://www.rosaceae.org/Analysis/15540493>); Rubus idaeus 'Joan J' genome (<https://www.rosaceae.org/Analysis/14031373>); Rosa rugosa genome (<https://www.rosaceae.org/Analysis/11775539>); Rosa chinensis 'Old Blush' genome (<https://www.rosaceae.org/analysis/282>); Rosa wichuraiana 'Basye's Thornless' genome (<https://www.rosaceae.org/Analysis/13087667>); Rosa chinensis 'Chilong Hanzhu' genome (<https://www.ncbi.nlm.nih.gov/bioproject/932466>; <https://doi.org/10.6084/m9.figshare.26888665.v1>).

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	<input type="checkbox"/> Not applicable
Reporting on race, ethnicity, or other socially relevant groupings	<input type="checkbox"/> Not applicable
Population characteristics	<input type="checkbox"/> Not applicable
Recruitment	<input type="checkbox"/> Not applicable
Ethics oversight	<input type="checkbox"/> 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	For population genomic analyses, a collection of 233 representative cultivated and wild rose accessions was used. Wild Rosa species include 10 sections, with comprehensive sampling covering all 10 sections, totaling 62 species. This study also includes 65 intermediate old cultivars and 42 modern cultivars, representing all types of roses in the current breeding history of Rosa.
Data exclusions	The North American group only contained three accessions, and thus was excluded from nucleotide diversity, linkage disequilibrium decay, and fixation index calculation.
Replication	RNA-seq experiments were performed in biological triplicates. Three distinct sample repetitions for Solanum lycopersicum cv. Heinz 1706 and Rosa hybrida 'Samantha®' were used in flow cytometry analysis. We confirm that all attempts at replication were successful.
Randomization	This study does not involve any randomized experimental group. All the sequencing libraries were prepared from multiple cells, so there is no need to randomize.
Blinding	Not applicable. Since our main results involved the assembly of a rose chromosome-level genome and the resequencing of rose accessions, blinding was not necessary.

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		Methods	
n/a	Involved in the study	n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies	<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines	<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology	<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern		
<input type="checkbox"/>	<input checked="" type="checkbox"/> Plants		

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
<input checked="" type="checkbox"/>	<input type="checkbox"/> Public health
<input checked="" type="checkbox"/>	<input type="checkbox"/> National security
<input checked="" type="checkbox"/>	<input type="checkbox"/> Crops and/or livestock
<input checked="" type="checkbox"/>	<input type="checkbox"/> Ecosystems
<input checked="" type="checkbox"/>	<input type="checkbox"/> Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

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

Plants

Seed stocks	The sampling locations for all plant materials are clarified in Supplementary Table 8.
Novel plant genotypes	Not applicable.
Authentication	Not applicable.

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	Leaves were placed in a 500 µL CyStain PI Absolute P Nuclei Extraction Buffer, chopped with a razor blade, and then filtered through a 50-µm filter. The collected cells were combined with 2000 µL of CyStain PI Absolute P Staining Buffer and incubated in the dark for 30 min.
Instrument	CyFlow Space Flow Cytometer (Sysmex Partec GmbH, Muenster, Germany).
Software	FCS Express™ Version 3 were used for data analysis.
Cell population abundance	Intact cells account for more than 85% of the total number of collected data.
Gating strategy	Total nuclei populations were gated using PI intensity. To ensure the settings are optimized for the nuclei population of interest, a selected region around them was drawn on the FSC vs SSC plot using the Polygon-Region tool. Firstly, pick as many particles as possible. Then pick the region with dense particles, try to remove adhesion particles and debris.

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