

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

Junping Gao

gao.jp@cau.edu.cn

China Agricultural University <https://orcid.org/0000-0002-9285-2539>

Zhao Zhang

China Agricultural University

Tuo Yang

China Agricultural University <https://orcid.org/0000-0001-5290-7454>

Yang Liu

China Agricultural University <https://orcid.org/0009-0001-7846-6025>

Shan Wu

Boyce Thompson Institute, Cornell University, Ithaca, NY, USA <https://orcid.org/0000-0001-8539-3808>

Honghe Sun

Boyce Thompson Institute; Plant Biology Section, School of Integrative Plant Science, Cornell University

Jie Wu

China Agricultural University

Yonghong Li

Shenzhen Polytechnic

Yi Zheng

Beijing University of Agriculture <https://orcid.org/0000-0002-8042-7770>

Haoran Ren

China Agricultural University

Yuyong Yang

Kunming Yang Chinese Rose Gardening Co., Ltd.

Shaochuan Shi

China Agricultural University

Wenyan Wang

China Agricultural University

Qi Pan

China Agricultural University

Lijuan Lian

Beijing Daxing Rose Garden

Shaowen Duan

Chinese Ancient Rose Garden

Yingxiong Zhu

Yunnan Xinhaihui Flowers Co., Ltd.

Youming Cai

Shanghai Academy of Agricultural Sciences

Hougao Zhou

Zhongkai University of Agriculture and Engineering

Hao Zhang

Flower Research Institute of Yunnan Academy of Agricultural Sciences

Kaixue Tang

Flower Research Institute of Yunnan Academy of Agricultural Sciences

Jiaopeng Cui

Beijing Botanical Garden

Dan Gao

Smartgenomics Technology Institute

Liyang Chen

Smartgenomics Technology Institute

Yunhe Jiang

China Agricultural University <https://orcid.org/0000-0002-7599-1422>

Xiaoming Sun

China Agricultural University

Xiaofeng Zhou

China Agricultural University <https://orcid.org/0000-0003-1239-7095>

Zhangjun Fei

Boyce Thompson Institute for Plant Research, Cornell University <https://orcid.org/0000-0001-9684-1450>

Nan Ma

China Agricultural University <https://orcid.org/0000-0002-2176-3133>

Article

Keywords:

Posted Date: July 16th, 2024

DOI: <https://doi.org/10.21203/rs.3.rs-4727628/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Version of Record: A version of this preprint was published at Nature Plants on October 11th, 2024. See the published version at <https://doi.org/10.1038/s41477-024-01820-x>.

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

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 Ren¹, Yuyong Yang⁷, Shaochuan Shi¹, Wenyan Wang¹, Qi Pan¹, Lijuan Lian⁸, Shaowen Duan⁹, Yingxiong Zhu¹⁰, Youming Cai¹¹, Hougao Zhou¹², Hao Zhang¹³, Kaixue Tang¹³, Jiaopeng Cui¹⁴, Dan Gao¹⁵, Liyang Chen¹⁵, Yunhe Jiang¹, Xiaoming Sun¹, Xiaofeng Zhou¹, Zhangjun Fei^{2,16*}, Nan Ma^{1*}, Junping Gao^{1*}

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 Applied Chemistry and Biotechnology, Shenzhen Polytechnic, 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

⁸Beijing Daxing Rose Garden, Beijing, China

⁹Chinese Ancient Rose Garden, Beijing, China

¹⁰Yunnan Xinhaihui Flowers Co., Ltd., Tonghai, Yunnan, China

¹¹Shanghai Academy of Agricultural Sciences, Shanghai, China

¹²College Horticulture & 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 100093, China

28 ¹⁵Smartgenomics Technology Institute, Tianjin 301700, China

29 ¹⁶USDA-ARS Robert W Holley Center for Agriculture and Health, Ithaca, NY

30 ¹⁷These authors contributed equally: Zhao Zhang, Tuo Yang, Yang Liu, Shan Wu, Honghe Sun, Jie
31 Wu, Yonghong Li.

32 *e-mail: gaojp@cau.edu.cn, zf25@cornell.com, ma_nan@cau.edu.cn

33

Abstract

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.

Introduction

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 vigor 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 East Asia, and *R. gallica* and *R. canina* throughout Europe².

Since the 18th 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, wherein 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 etc. 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 the good traits from Chinese and European old roses, has since become recognized as the class of modern rose (referred to as *Rosa hybrida*). The modern rose shows great phenotypic variation, as well as genetic potential for yield, quality, phenological adaptability, making it quickly predominate in rose cultivation^{5,6}. Nowadays, over 35,000 modern rose cultivars, of which most are tetraploid, have been cultivated worldwide. These modern roses have a wide variety of colors, flower shapes and scent, as well as different plant architectures, all of which are important traits and need to be better understood.

According to the 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

resulted from the reticulate evolution, a high percentage of repetitive sequences, and the polyploidy all challenge the deciphering of 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 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-centered blooms, continuous and recurrent flowering traits, and leathery foliage (Supplementary Fig. 1), and 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 two hundred 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 for the tetraploid modern rose

To develop a reference genome for tetraploid modern roses, we sequenced the genome of cultivar ‘Samantha®’ ($2n = 4x = 28$) (Supplementary Fig. 1), which had an estimated tetraploid genome size of approximately 1.91-2.13 Gb based on *k*-mer and flow cytometry analyses (Supplementary Figs. 2 and 3 and Supplementary 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 chromosome 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 of 37.76 Mb. The Hi-C and Pore-C data were used to further assemble 91.6% of the

contigs (1,971 Mb) into 28 pseudochromosomes (Fig. 1 and Supplementary Table 3). The homologous chromosomes of ‘Samantha®’ were named based on their potential relative relationship with 17 wild species distributed in China or Chinese old roses (Supplementary Note 1).

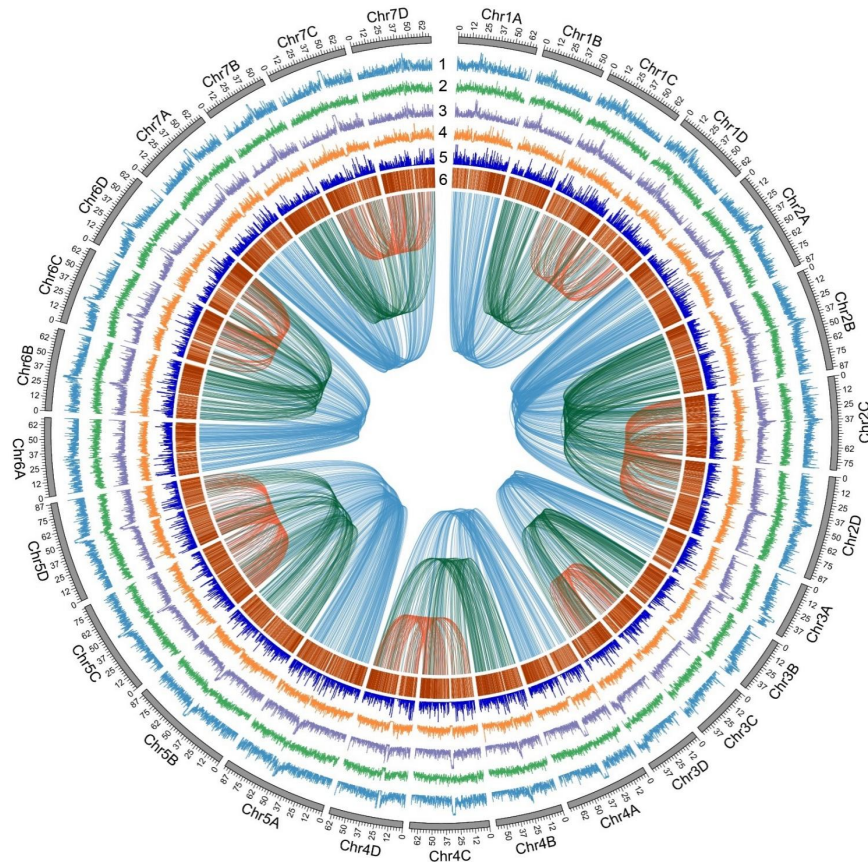


Fig. 1 Genomic landscape of *R. hybrida* ‘Samantha®’. 1-6, LTR-RT density (1), *Copia*-type LTR-RT density (2), *Gypsy*-type LTR-RT density (3), DNA transposon density (4), gene density (5), and gene expression levels (6) in 500 kb non-overlapping windows. Syntenic blocks between homologous chromosomes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

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. Additionally, around 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. 4). 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. 5), 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. 6) and the high degree of synteny with the genome of *R. chinensis* ‘Old Blush’⁸ (Supplementary Fig. 7). Both Pore-C and Hi-C signal heatmaps supported the high continuity and phasing accuracy of the assembly (Supplementary Fig. 8). 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. 9). 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 2-3), 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 analyzed 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*, *Banksianae*, *Bracteatae*, and *Laevigatae*), 65 intermediate old cultivars, and 42

modern cultivars. In total, we identified 17,567,247 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 nonsynonymous 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 of those 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 the kinship analysis for the 233 rose accessions using the KING program¹⁴ and did not identify any duplicated accessions or 1st-degree relatives in this rose collection. We then evaluated the nucleotide diversity of different rose populations. The nucleotide diversity (π) of six different groups (section Chinenses, section Rosa, section Synstylae, section Cinnamomeae, section 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 to that of the intermediate old cultivars, but significantly higher than any other groups (Supplementary Table 11). This suggests a weak bottleneck effect during the domestication of modern roses and indicates that modern roses still have great genetic potential for improvement.

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, though 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,

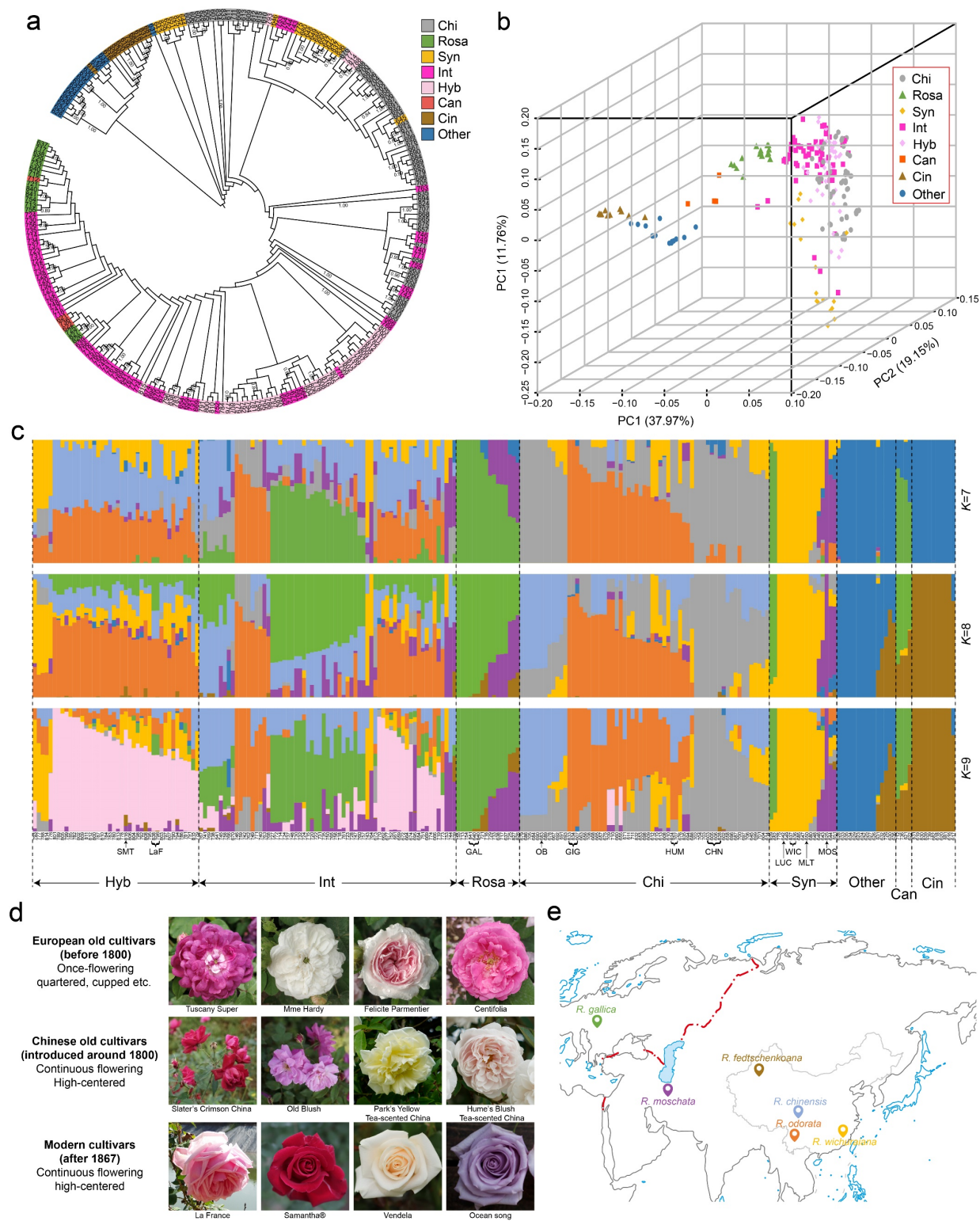


Fig. 2 Population structure of 233 accessions within genus *Rosa*. a, Phylogenetic tree constructed using SNPs at fourfold synonymous third-codon transversion (4DTv) sites. Hyb, modern cultivars, *R. hybrida*; Int, intermediate old cultivars; Syn, section Synstylae; Chi, section Chinenses; Rosa, section Rosa; Can,

section Caninae; Cin, section Cinnamomeae; Other, sections Pimpinellifoliae, Microphyllae, Bracteatae, and Banksianae. **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 color 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. wichuraiana*; 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 of flower shape along the rose breeding history. **e**, Geographical distribution of the six original species. *R. chinensis* is native to Zhejiang, Guangdong, Guangxi, Fujian, and Taiwan in China. *R. wichuraiana* is native to Hubei, Sichuan, and Guizhou in China. *R. gallica* is native to central and southern Europe and Western Asia. *R. fedtschenkoana* is native to Xinjiang in China. *R. odorata* is native to Yunnan in China. *R. moschata* is believed to have originated from South Asia, including Iran, Afghanistan, Pakistan, and India.

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. Consistently in the principal component analysis (PCA), accessions from section Cinnamomeae and other sections were 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 variation after human selection along 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 18th-19th 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. 10). 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 the *R. chinensis* and the other consisting of the *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 harbored 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 favored 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) had flowers in large clusters or trusses and 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 (Supplementary Fig. 11). 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), and thus the exact contribution of each species to this inflorescence trait of modern roses still remained unclear.

Before the introduction of Chinese germplasms around 1800, European old cultivars, e.g., Gallica, Damask, Alba, and Centifolia (top panel in Fig. 2d), 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 (middle panel in Fig. 2d). 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 color 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 laid the foundation of the favored flower shape of modern cultivars (bottom panel in Fig. 2d). It is worth mentioning that *R. odorata*-derived old cultivars exhibit a high-centered 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 (bottom panel in Fig. 2d) mimicked that of *R. odorata*-derived old cultivars, while with increased petal numbers. Collectively, *R. odorata* and derived cultivars contributed a larger portion to the genome structure and favored 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). In order to better analyze the genomic composition of *R. hybrida* ‘Samantha®’, the origins of different regions of each chromosome were inferred based on 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 Supplementary Fig. 12). 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 chimerism, each derived from multiple ancestral species except Chr7B, indicating extensive homeologous exchanges and/or introgressions in the ‘Samantha®’ genome (Supplementary Fig. 12) 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 homeologous exchanges. Furthermore, a large inversion between the homologous chromosomes was observed on chromosome 7 (Fig. 3d), which was further supported by the local Hi-C heatmaps (Supplementary Fig. 13) and the collinearity with the four published *Rosa* genomes^{8,16-18} (Supplementary Fig. 14). Genomic regions of Chr7A and Chr7C harboring 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.

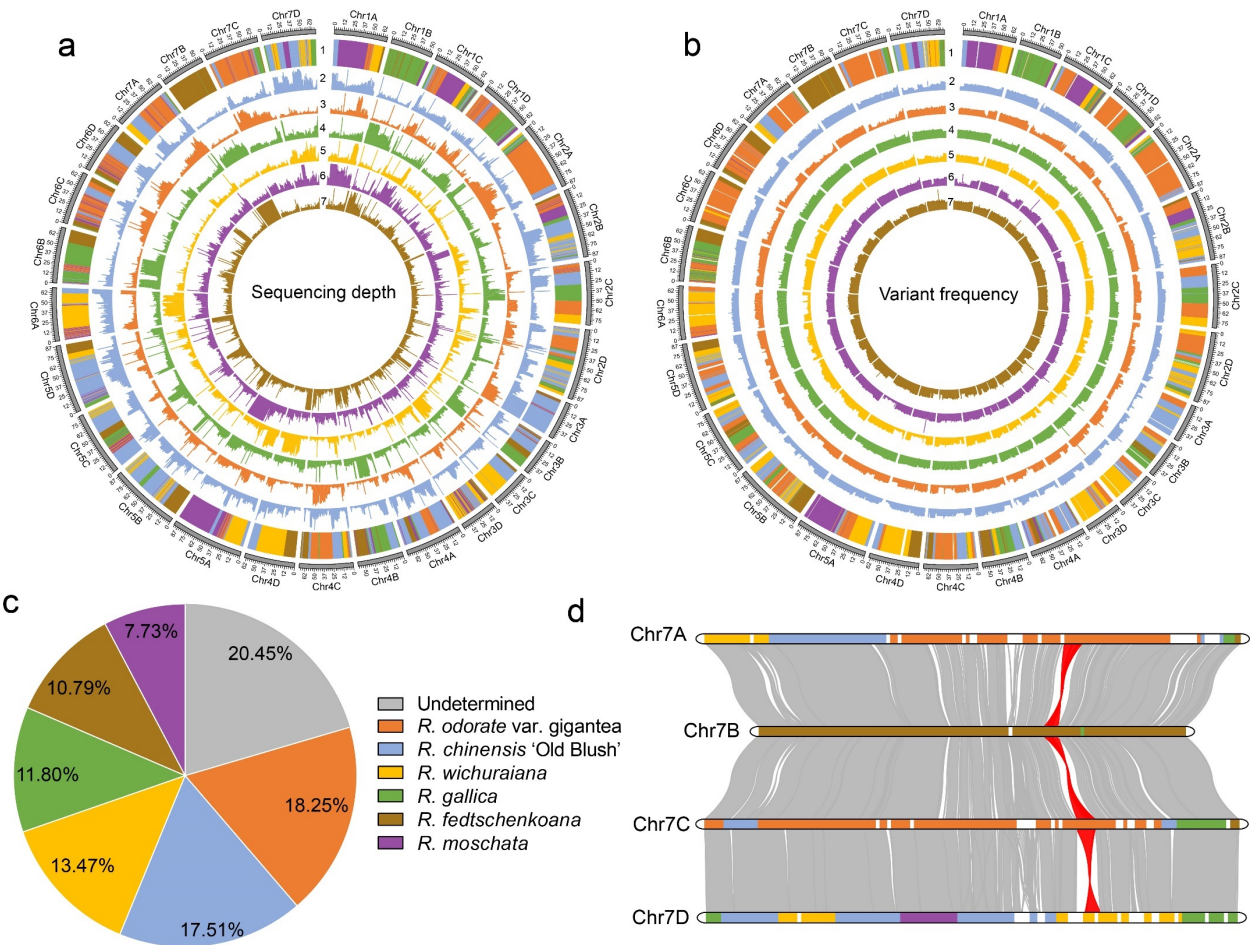


Fig. 3 Genetic compositions of the ‘Samantha®’ chromosomes. **a**, Origins of different regions on the chromosomes (1) inferred by the depth coverage of reads from wild ancestors, *R. chinensis* ‘Old Blush’ (2), *R. odorata* var. *gigantea* (3), *R. gallica* (4), *R. wichuraiana* (5), *R. moschata* (6), 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. Colors 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.

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 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 that of Hyb_4 ($\pi = 4.83 \times 10^{-3}$), reflecting the function of human selection during cultivar improvement (Fig. 4a). Additionally, Int_4 exhibited a faster LD decay than Hyb_4 (Fig. 4b), again supporting the decreased genetic diversity for recently formed modern cultivars. Through comparing Int_4 with Hyb_4 and using combined filtering of π ratio and 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 color, senescence, growth, and disease resistance were revealed (Fig. 4c).

The ability of continuous and recurrent flowering is among the most important traits of modern roses. It is the basis of cut rose production and also highly valued for garden/potted roses. This character is believed to be inherited from *R. chinensis*^{2,8}. The *Koushin* (*KSN*) gene in rose has been found to be involved in the determination of once/continuous flowering^{19,20}. Here, the *KSN* genes (*Rh3AG182500*; *Rh3BG208800*) were found in a selective sweep region (Supplementary Fig. 15). Clustering analysis of SNP genotype in this selective sweep region further supported the contribution of section *Chinenses* to the continuous and recurrent flowering of modern roses (Supplementary Fig. 15). Furthermore, genes encoding regulators involved in vernalization²¹ (e.g.,

VERNALIZATION INSENSITIVE 3-like 1, *VERNALIZATION INDEPENDENCE*, *FRIGIDA-like*, *RELATED TO VERNALIZATION1*) and the photoperiod pathway^{22,23} (e.g., *NO FLOWERING IN SHORT DAY*, *LATE FLOWERING*) were also intensively selected, suggesting that the trait of continuous and recurrent flowering may have been realized by fixing genes in separate flowering pathways.

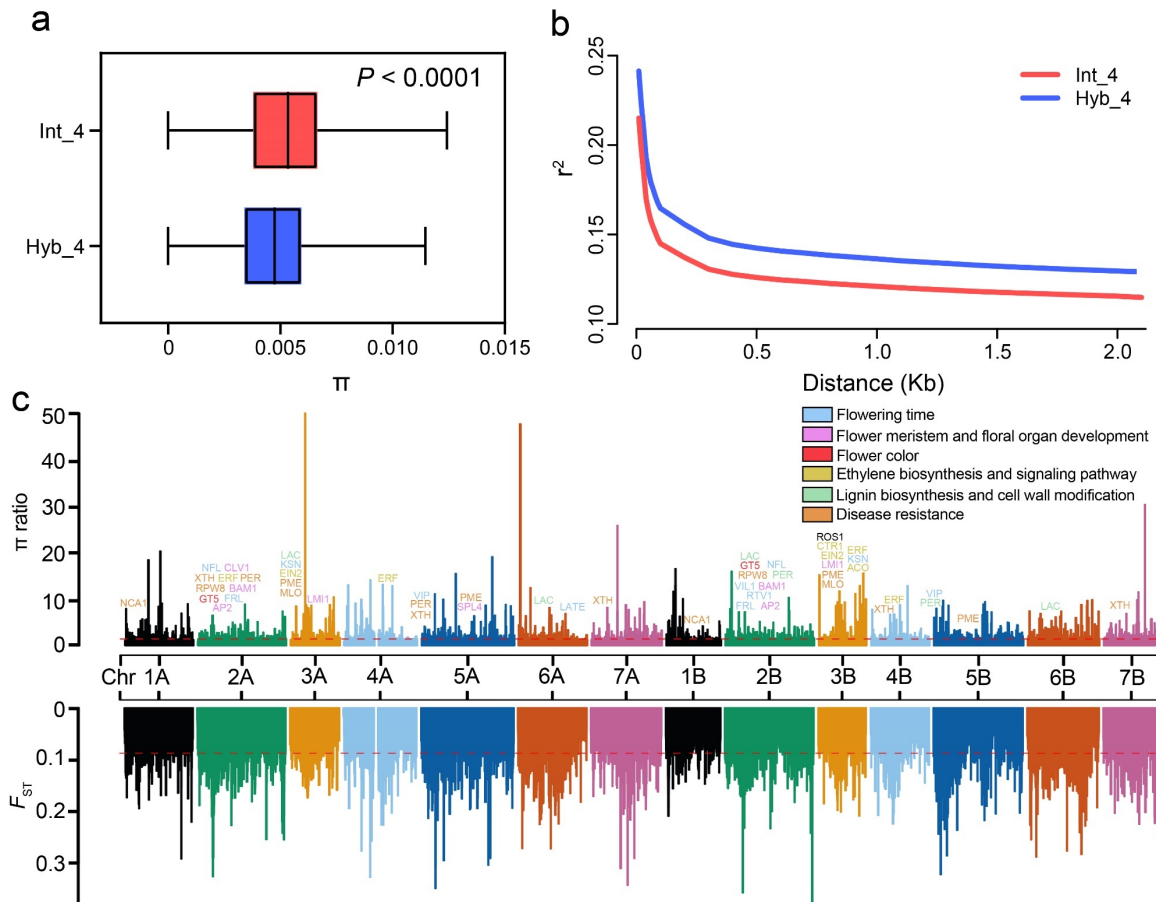


Fig. 4 Genome-wide distribution of selective sweeps in *R. hybrida*. **a**, Nucleotide diversity (π) of tetraploid intermediate old cultivars (Int_4; $n = 45$) and tetraploid modern rose cultivars (Hyb_4; $n = 32$). For each box plot, the lower and upper bounds of the box indicate the first and third quartiles, respectively, and the center line indicates the median. P -value was calculated using a two-sided Student's t -test. **b**, LD 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 indicated the top 5% cutoff values. Gene abbreviations: NCA1, NO CATALASE ACTIVITY 1; KSN, Koushin; CLV1, CLAVATA 1; NFL, NO FLOWERING IN SHORT DAY; FRL, FRIGIDA-like; VIL1, VERNALIZATION INSENSITIVE 3-like 1; SPL4, squamosa promoter binding protein-like 4; LM1, LATE MERISTEM IDENTITY1; LATE, LATE FLOWERING; VIP, VERNALIZATION INDEPENDENCE; GT5, UDP-glucose flavonoid 3-O-glucosyltransferase 5; RTV1, RELATED TO VERNALIZATION 1; AP2, APETALA 2; CTR1,

CONSTITUTIVE TRIPLE RESPONSE 1; *ACO*, 1-aminocyclopropane-1-carboxylate oxidase; *EIN2*, *ETHYLENE INSENSITIVE 2*; *ERF*, ethylene response factor; *LAC*, LACCASE; *PER*, PEROXIDASE; *PME*, pectinesterase; *XTH*, xyloglucan endotransglucosylase/hydrolase; *RPW8*, resistance to powdery mildew 8; *MLO*, mildew resistance locus O; *ROS1*, REPRESSOR OF SILENCING1.

Flower meristem identity genes *SPL4* (*SQUAMOSA PROMOTER BINDING PROTEIN-LIKE 4*)²⁴, *CLAVATA 1* (*CLV1*)²⁵, *BARELY ANY MERISTEM 1* (*BAM1*)²⁶ and *LATE MERISTEM IDENTITY1* (*LM1*)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 along modern rose breeding. In rose, floral identity gene *APETALA 2* (*AP2*) have been proven to regulate petal numbers through stamen petaloidy²⁸⁻³⁰. Here, we identified that *AP2* genes (*Rh2AG184600*; *Rh2BG195000*) have been under selection during modern rose breeding (Supplementary Fig. 16). We performed clustering on the SNP genotypes of these two selective sweep regions that contained *AP2* genes, respectively, and found that SNP profiles of modern cultivars were highly similar to those of section *Chinenses* (Supplementary Fig. 16). This directly demonstrates the crucial contribution of section *Chinenses* to the petal number in modern roses. Moreover, petal color is among the most attractive characteristics of modern roses. *UDP-glucose flavonoid 3-O-glucosyltransferase 5* involved in floral pigment biosynthesis³¹ were 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 play a pivotal role in governing flower opening and senescence in rose³²⁻³⁴. Ethylene-related genes were found in several highly selected regions, including biosynthesis-related genes (*ACOs*) and genes encoding signaling components (*CTR1*, *EIN2*) and ethylene-responsive transcription factors (*ERFs*), indicating that both ethylene biosynthesis and signaling pathways have been selected during the breeding history of modern roses. *CTR1*³⁵ (*Rh3BG199500*), the key component of the ethylene signal transduction pathway, has been under selection, and its SNP genotype profile was similar to that of *R. chinensis* (Supplementary Fig. 17). Furthermore, *ROS1*³⁶ (*Rh3BG199400*), encoding a DNA demethylase,

was also found in this selective sweep region (Supplementary Fig. 17), indicating that DNA methylation might play a critical role in determining rose traits during its breeding.

Additionally, genes encoding enzymes related to lignin biosynthesis and cell wall modification, e.g., *xyloglucan endotransglucosylase/hydrolase*³⁷ (*XTH*), *LACCASE* (*LAC*), *PEROXIDASE* (*PER*), and *pectinesterase* (*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* are 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 to intermediate old cultivars, disease resistance in modern cultivars has been selected during the last two hundred years of modern rose breeding history. In the selective sweep regions, genes homologous to *Arabidopsis NCA1*⁴³ that play a crucial role in the clearance of hydrogen peroxide, *Arabidopsis RPW8*⁴⁴ involved in powdery mildew resistance, and rose *mildew resistance locus O* (*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 gene further confirm the importance of disease resistance in breeding of modern cultivars.

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 role⁴⁷. We were able to identify the *OOMT1* gene (*Rh2AG330600*) gene located in a region on chromosome Chr2A of the ‘Samantha®’ genome that is derived from *R. odorata* var. *gigantea*. No *OOMT1* genes were identified in other homologous chromosomes of ‘Samantha®’. While the *OOMT1* gene was not identified to be under selection, a 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 roses⁴⁷. 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 domestication and breeding of 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 consumers' preference and our genotype analysis of the *OOMT1* gene provide 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 eastern and western 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 significance for understanding its origin and domestication, 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 domestication of roses. The genome of 'Samantha®' is one of the few **phased chromosome-scale** polyploid plant genome assemblies reported to date, 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 the creation of modern rose through hybridizations in the 19th 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 to *R. chinensis* ‘Old Blush’.

Over thousands of years, the cultivation and domestication of old cultivars in both China and Europe have accumulated a significant 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, fragrance, 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 in order to delay senescence. By controlling the crucial component upstream of the ethylene signaling 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 the miRNA172 cleavage, promoting the double flower phenotype²⁸, and transposon insertion in the *KSN* gene restricts its expression, leading to the continuous and recurrent flowering trait of modern roses⁴⁸. In ‘Samantha®’, both *AP2L* and *KSN* harbor 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, e.g., those that have been selected during breeding based on 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) kit and sequenced on the PromethION. PacBio HiFi libraries with inserts of ~20 kb were prepared following the PacBio HiFi library construction protocol, and sequenced on the PacBio Revio system. For MGI sequencing, a DNA sequencing library was constructed with an insert size of ~300 bp and sequenced on an DNBSEQ-T7 platform. The Pore-C libraries were generated according to the method described in Li et al.⁵⁰. Briefly, approximately 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 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) and centrifuged at 2,200 g 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 ZYMO DNA Clean & Concentrator kit (ZYMO, #D4014). DNA fragments larger than 1 kb were selected and sequenced using the Nanopore Ligation Sequencing Kit (SQK-LSK109). Hi-C libraries were prepared following the method described in Belton et al.⁵¹ with minor modifications. Briefly, calluses were vacuum infiltrated in nuclei isolation buffer supplemented with 2% formaldehyde for 15 min and then ground with liquid nitrogen and washed orderly by 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 applied with restriction enzyme (400 units DpnII) at 37 °C on rocking platform. The constructed Hi-C libraries were sequenced on an DNBSEQ-T7 platform.

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 was constructed using the Illumina Genomic DNA Library Prep Kit following manufacturer's recommendations and sequenced on an Illumina NovaSeq 6000 platform.

Genome size estimation and ploidy analysis using flow cytometry

The genome size of 'Samantha®' was estimated using flow cytometry⁵². 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 2,000 µL of CyStain PI Absolute P Staining Buffer and incubated in the dark for 30 min. The nuclei suspension was

analyzed using CyFlow Space Flow Cytometer (Sysmex Partec). Genome size estimations were independently determined using *Solanum lycopersicum* cv. Heinz 1706 as the internal reference standards.

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 analyzed using CyFlow Ploidy Analyser (Supplementary Fig. 18).

RNA extraction and transcriptome sequencing

Total RNA from floral buds, fruits, leaves, petals, pistils and stamens, roots, seeds, stems, and thorns of ‘Samantha®’ was extracted using the hot borate method as previously described³⁴. Strand-specific RNA-Seq libraries were constructed using the protocol described in Zhong et al.⁵³ and sequenced on the DNBSEQ-T7 platform. 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 hifiasm⁵⁴ (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 common contaminants database (ftp://ftp.ncbi.nlm.nih.gov/pub/kitts/contam_in_euks.fa.gz) and nt database (<ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nt.gz>) using BLASTN⁵⁵ (v.2.12.0+) with an e-value cutoff of $1e-5$. 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. Firstly, the DpnII restriction sites on the contigs were analyzed using Cooler⁵⁶ (v.0.8.6). Subsequently, Pore-C reads were aligned to the contigs using minimap2⁵⁷ (v.2.17-r941), and fragment annotation and generation of the pairwise contact matrix were carried out using the HiPore-C pipeline (<https://github.com/zhengdafangyuan/HiPore-C>).

C). The pairwise contact matrix was used to anchor contigs using LACHESIS⁵⁸. To identify potential collapsed regions in the assembled chromosomes, PacBio HiFi reads were aligned to the assembled genome using minimap2⁵⁷ (v.2.17-r941) with default parameters. Based on the alignments the mapping depth was calculated for each 100-kb non-overlapping window across the assembly using mosdepth⁵⁹. Genome regions with a depth greater than 100× (twice the normal depth of 50×) are identified as potential collapsed regions. The assembled contigs were then aligned to these regions using minimap2⁵⁷ (v.2.17-r941) to obtain the corresponding contigs, which were then re-anchored to the chromosomes using HiCUP⁶⁰ (v.0.6.1). The final haplotype-resolved chromosome assembly results were manually corrected using JuiceBox⁶¹ (v.1.11.08) based on 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 nPhase¹³ (v.1.2.0). Synteny between the assembly and the *R. chinensis* ‘Old Blush’ genome⁸ was identified using SYRI⁶² (v.1.7.0), and synteny between the assembly and the K5 genetic map of modern rose¹² was identified using ALLMAPS⁶³ (v.0.8.12). BUSCO¹⁰ (v.5.4.5) 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 program⁶⁴ (v.2.4.2). Merqury⁶⁵ (v.1.3) was further utilized to assess the consensus quality value and *k*-mer completeness of the assembly.

Repeat sequence 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 LTRharvest⁶⁶ (v.1.5.10) and MITE-Hunter⁶⁷ (v.11-2011), respectively. The ‘Samantha®’ assembly was then masked using RepeatMasker⁶⁸ (v.4.0.6) with the LTR-RT and MITE libraries. The unmasked sequences in the genome were further searched for repeat elements using RepeatModeler⁶⁹ (v.1.0.11). All the repetitive sequences generated above were combined into one *de novo* repeat library, and along with the Repbase⁷⁰ (v.29.04) and Dfam⁷¹ (v.3.8) databases, were used to search the genome for repeat sequences using RepeatMasker⁶⁸ (v.4.0.6).

The repeat-masked genome was used for protein-coding gene prediction with the EVidenceModeler pipeline⁷² (v.2.1.0), which combines evidence of *ab initio* gene prediction, transcript mapping, and protein homology to define the final gene models. SNAP⁷³ (v.2006-07-28), GlimmerHMM⁷⁴ (v.3.0.4) and AUGUSTUS⁷⁵ (v.3.3) were used for *ab initio* gene predictions. The paired-end RNA-Seq data were cleaned with Trimmomatic⁷⁶ (v.0.39) and assembled using Trinity⁷⁷ (v.2.4.0) and PASA⁷⁸ (v2.5.3), with the *de novo* mode and the genome-guided mode, respectively. The assembled transcripts were used as the transcript evidence. Protein sequences from *Rubus*⁷⁹, *Malus*⁸⁰, *Fragaria*⁸¹ and diploid rose^{8,16-18}, were aligned to the ‘Samantha®’ genome using miniprot⁸² (v.0.13) to provide the protein homology evidence. For gene function annotation, protein sequences of the predicted genes were compared against *Arabidopsis* protein⁸³ and UniProt⁸⁴ (Swiss-Prot/TrEMBL) (v.2023_01) databases using BLAST, as well as the InterPro database using InterProScan⁸⁵ (v.5.0). GO annotations were obtained using Blast2GO⁸⁶ (v.6.0). NLR-Annotator⁸⁷ (v.2.1b) is 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 Trimmomatic⁷⁶ (v.0.39). ChrA and ChrB were selected to represent the ‘Samantha®’ genome for variation calling. The cleaned reads were aligned to ChrA and ChrB, respectively, using BWA-MEM⁸⁸ (v.0.7.17) with default parameters. Picard (v2.7.1; <http://broadinstitute.github.io/picard/>) was used to mark duplicated alignments. The HaplotypeCaller function in GATK⁸⁹ (v.4.2.6) was then used to generate a GVCF file for each of the 233 rose accessions. All GVCF files were combined for variant calling using the function GenotypeGVCFs in GATK⁸⁹. Hard filter was applied to the identified raw SNPs using GATK with 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) greater than 0.05 and a data missing rate less than 10%. For subsequent SNP analysis, the scoring was based on codominant markers. SNP

annotation was performed using the package ANNOVAR⁹⁰ (v.2015-12-14).

Phylogenetic and population genomic analysis

A total of 501,064 SNPs at the 4DTv (four-fold synonymous third-codon transversion) sites were used for phylogenetic and population structure analyses. Maximum-likelihood phylogenetic tree was constructed using FastTree⁹¹ (v. 2.1.11) with 1,000 bootstraps. PCA was performed using PLINK⁹² (v.1.90b6.10). Population structure was investigated using ADMIXTURE⁹³ (v.1.3.0) 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 CV error, which indicates 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 VCFtools⁹⁴ (v.0.1.16). Fixation index (F_{ST}) values were calculated between two groups (tetraploid intermediate old cultivars and tetraploid modern cultivars) using VCFtools⁹⁴ (v.0.1.16).

LD decay patterns were calculated using PopLDdecay⁹⁵ (v.3.42) with the default parameters for the two groups (tetraploid intermediate old cultivars and tetraploid modern cultivars). Correlation coefficient (r^2) of genotypes was used as a measure of the LD level. SNPs within each group were extracted for the analysis. SNPs within 2-kb sliding windows were used to estimate average r^2 at various physical distance classes, and the LD 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 based on 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-TREE⁹⁶ (v.1.6.8), with a bootstrap value of 1,000. The genotypes of SNPs in each selective sweep region were presented through a heatmap, which was generated using 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 amount 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, if the highest sequencing coverage depth among the six species was 1.5 times higher than the second highest, this 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 based on 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:

$$\text{Variant frequency} = \frac{\text{Number of homozygous variant sites} + \frac{1}{2} \times \text{Number of heterozygote sites}}{\text{Total number of identified sites}}$$

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®’,

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.

Data availability

Raw reads generated in this study have been deposited in NCBI BioProject database under accession numbers PRJNA1108167 and PRJNA704782. The sequences and annotations of the ‘Samantha®’ genome assembly are available at Figshare (<https://doi.org/10.6084/m9.figshare.22774097>). The reviewer’s links to the submitted datasets are provided below:

PRJNA1108167: <https://dataview.ncbi.nlm.nih.gov/object/PRJNA1108167?reviewer=q3ii6fdakvfme19pulek65vja2>

PRJNA704782: <https://dataview.ncbi.nlm.nih.gov/object/PRJNA704782?reviewer=s1okd75jhm2q6roocnbsnvr5>

Figshare: <https://figshare.com/s/b31a79f7114120be19cf>

Acknowledgements

We thank Dr. Tao Lin (China Agricultural University) and Dr. Qiang Gao (BGI Tech., Beijing) for helpful discussions, Dr. Haibo Xin and Dr. Yanhua 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 the help of flow cytometry experiments, and Wuhan Grandomics Biosciences Co., Ltd. for providing Pore-C data 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.), National Key Research and Development Program of China (Grant no. 2018YFD1000400 to J.G.), General Project of Shenzhen Science and Technology and Innovation Commission (Grant no. 21K270360620 to

Y.Li.), National Natural Science Foundation of China (Grant no. 31572162 to Y.Li., 31772344 and 31972444 to Z.Z., and 31522049 and 31872148 to N.M.), Postdoctoral Fellowship Program of CPSF (Grant no. GZB20240831 to T.Y.).

Author contributions

J.G., N.M. and Z.F. 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.Z. integrated the genome assembly and annotation. T.Y., H.S. performed selective sweep analysis. T.Y., H.S., J.W., and Y.Z. conducted phylogenetic and population genomic analysis. T.Y., Y. Liu., J.W., S.W., Z.Z., Y.L., Z.F., N.M., and J.G. wrote and revised the manuscript.

Competing interests

The authors declare no competing interests.

Figure legends

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 center.

Fig. 2 Population structure of 233 accessions within genus *Rosa*. **a**, Phylogenetic tree constructed using SNPs at fourfold 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, and Banksianae. **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 color 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. wichuraiana*; 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 of flower shape along the rose breeding history. **e**, Geographical distribution of the six original species. *R. chinensis* is native to Zhejiang, Guangdong, Guangxi, Fujian, and Taiwan in China. *R. wichuraiana* is native to Hubei, Sichuan, and Guizhou in China. *R. gallica* is native to central and southern Europe and Western Asia. *R. fedtschenkoana* is native to Xinjiang in China. *R. odorata* is native to Yunnan in China. *R. moschata* is believed to have originated from South Asia, including Iran, Afghanistan, Pakistan, and India.

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. Colors 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.

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 each box plot, the lower and upper bounds of the box indicate the first and third quartiles, respectively, and the center line indicates the median. *P*-value was calculated using a two-sided Student’s *t*-test. **b**, LD 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 indicated the top 5% cutoff values. Gene abbreviations: *NCA1*, NO CATALASE ACTIVITY 1; *KSN*, Koushin; *CLV1*, CLAVATA 1; *NFL*, NO FLOWERING IN SHORT DAY; *FRL*, FRIGIDA-like; *VIL1*, VERNALIZATION INSENSITIVE 3-like 1; *SPL4*, squamosa promoter binding protein-like 4; *LMII*, LATE MERISTEM IDENTITY1; LATE, LATE FLOWERING; *VIP*, VERNALIZATION INDEPENDENCE; *GT5*, UDP-glucose flavonoid 3-O-glucosyltransferase 5; *RTV1*, RELATED TO VERNALIZATION 1; *AP2*, APETALA 2; *CTR1*, CONSTITUTIVE TRIPLE RESPONSE 1; *ACO*, l-aminocyclopropane-1-carboxylate oxidase; *EIN2*, ETHYLENE INSENSITIVE 2; *ERF*, ethylene response factor; *LAC*, LACCASE; *PER*, PEROXIDASE; *PME*, pectinesterase; *XTH*, xyloglucan endotransglucosylase/hydrolase; *RPW8*, resistance to powdery mildew 8; *MLO*, mildew resistance locus O; *ROS1*, REPRESSOR OF SILENCING1.

References

1. 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).
2. Roberts, A. V., Debener, T. & Gudin, S. *Encyclopedia of Rose Science*. Academic Press: Cambridge, MA, US (2003).
3. Yokoya, K., Roberts, A., Mottley, J., Lewis, R. & Brandham, P. Nuclear DNA amounts in roses. *Ann. Bot.* **85**, 557–561 (2000).
4. Wylie, A. P. The history of garden roses. *J. Roy. Hort. Soc.* **79**, 555–571 (1954).
5. de Vries, D. & Dubois, L. Rose breeding: past, present, prospects. *Acta Hort.* **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. Hort.* **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 prickles: genomic insights linked to moisture adaptation. *Natl. Sci. Rev.* **8**, nwab092 (2021).
19. Randoux, M. *et al.* Gibberellins regulate the transcription of the continuous flowering regulator, *RoKSN*, a rose *TFL1* homologue. *J. Exp. Bot.* **63**, 6543–6554 (2012).
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. Amer. Soc. Hort. 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. Scalliet, G. *et al.* Scent evolution in Chinese roses. *Proc. Natl. Acad. Sci. USA* **105**, 5927–5932 (2008).
48. Bai, M. *et al.* KSN heterozygosity is associated with continuous flowering of *Rosa rugosa* Purple branch. *Hortic. Res.* **8**, 26 (2021).
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. Zhong, S. *et al.* High-throughput Illumina strand-specific RNA sequencing library preparation. *Cold Spring Harb. Protoc.* **2011**, 940–949 (2011).
54. 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).
55. Boratyn, G. M. *et al.* Domain enhanced lookup time accelerated BLAST. *Biol. Direct* **7**, 12 (2012).
56. Abdennur, N. & Mirny, L. A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* **36**, 311–316 (2019).
57. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

58. Burton, J. N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
59. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
60. Wingett, S. W. *et al.* HiCUP: pipeline for mapping and processing Hi-C data. *FI000Res.* **4**, (2015).
61. Durand, N. C. *et al.* Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
62. 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).
63. Tang, H. *et al.* ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol.* **16**, 1–15 (2015).
64. 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).
65. 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).
66. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
67. 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).
68. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **25**, Chapter 4:Unit 4.10 (2009).
69. Flynn, J.M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
70. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
71. Hubley, R. *et al.* The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**, D81–D89 (2015).
72. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
73. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
74. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
75. 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).
76. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
77. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).

78. Haas, B. J. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
79. Zhou, J. *et al.* Comparison of red raspberry and wild strawberry fruits reveals mechanisms of fruit type specification. *Plant Physiol.* **193**, 1016–1035 (2023).
80. 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).
81. Mao, J. *et al.* High-quality haplotype-resolved genome assembly of cultivated octoploid strawberry. *Hortic. Res.* **10**, uhad002 (2023).
82. Li, H. Protein-to-genome alignment with minimap. *Bioinformatics* **39**, btad014 (2023).
83. Cheng, C.-Y. *et al.* Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J.* **89**, 789–804 (2017).
84. Bairoch, A. The universal protein resource (UniProt). *Nucleic Acids Res.* **33**, D154–D159 (2005).
85. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
86. Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
87. Steuernagel, B. *et al.* The NLR-Annotator tool enables annotation of the intracellular immune receptor repertoire. *Plant Physiol.* **183**, 468–482 (2020).
88. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
89. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
90. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
91. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
92. 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).
93. 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).
94. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
95. 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).
96. 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).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [RoseSuppNoteFigureR1.pdf](#)