Haplotype-resolved genome assembly provides insights into the floral scent of *Rosa rugosa*

Xi Cheng
Institute of Grassland, Flowers and Ecology, Beijing Academy of Agriculture and Forestry Sciences
Dan Gao
Smartgenomics Technology Institute
Hongli Wang
Institute of Grassland, Flowers and Ecology, Beijing Academy of Agriculture and Forestry Sciences
Guoliang Wang
Beijing Key Laboratory of Agricultural Genetic Resources and Biotechnology, Institute of Biotechnology, Beijing Academy of Agriculture and Forestry Sciences
Dongliang Chen
Institute of Grassland, Flowers and Ecology, Beijing Academy of Agriculture and Forestry Sciences
Chang Luo
Institute of Grassland, Flowers and Ecology, Beijing Academy of Agriculture and Forestry Sciences
Hua Liu
Institute of Grassland, Flowers and Ecology, Beijing Academy of Agriculture and Forestry Sciences
Tianyi Wang
Smartgenomics Technology Institute
Chengzhi Jiao
Smartgenomics Technology Institute
Kezhong Zhang
College of Landscape, Beijing University of Agriculture
Bei-bei Jiang Jiang
Sichuan Agricultural University
Conglin Huang (conglinh@126.com)
Institute of Grassland, Flowers and Ecology, Beijing Academy of Agriculture and Forestry Sciences

Article

Keywords: Haplotype genome, Rosa rugosa, floral scent, 2-phenylethanol, allele-specific expression

Posted Date: April 3rd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2757151/v1

Abstract

Rose is an important aromatic plant and produces flowers that are used in medicine and food. We herein present a haplotype-resolved genome for *Rosa rugosa*cultivar Hanxiang. Analyses of allele-specific expression identified a potential mechanism underlying floral scent biosynthesis. Population genomic analyses involving 133 *Rosa* accessions elucidated evolutionary histories and a single *R. rugosa* domestication event. Pathways mediating the synthesis of scent-related metabolites were enriched according to the analyses of the transcriptomes, haplotype variations, and allelic imbalances during the flower development stages of Hanxiang and Guomeigui (*R. rugosa* accessions with diverse fragrances). The enzyme-encoding ASE genes *RrHX1G119800* and *RrHX1G204700* (primary amine oxidases) and *RrHX2G284700* (L-tryptophan decarboxylase) in the phenylethylamine pathway were tentatively designated as core genes useful for improving 2-phenylethanol production in rose flowers. Our results provide molecular insights into the formation of *R. rugosa* floral fragrances and genome-level data that are useful for enhancing rose traits via genetic engineering.

Introduction

Floral volatiles are biologically and economically important compounds that are vital for attracting pollinators, while also contributing to plant defense responses and interactions with the environment[1]. They can also considerably affect human health. Current studies on plant genomes have clarified the genetic basis of the internal quality-related properties of representative *Rosa* species, with implications for enhancing floral scents[2,3].

Rosa rugosa, a deciduous shrub belonging to the family Rosaceae, originated in China and produces flowers that are used in medicine and food. Additionally, *R. rugosa* is a rich source of amino acids and trace elements needed by humans and may be useful for treating endocrine disorders and promoting blood circulation[4]. Its petals contain relatively high amounts of terpenes, phenols, esters, and alcohols, including phenylethanol, citronellol, geraniol, linalool, and eugenol, which have neuroprotective[5], antioxidative[6,7], anxiolytic[8], anti-inflammatory[9], and skin-protective[10] effects. *Rosa rugosa* has a long history as an important floral species in China. It is a commonly cultivated flower species with characteristics that make it useful in many industries. Long-term artificial or natural selection has resulted in highly diverse *R. rugosa* germplasm resources, which are mainly distributed in northwestern, southwestern, and northern China and in other regions in Asia (e.g., Japan and South Korea) and elsewhere.

Because of its highly heterozygous genome, *R. rugosa* is a valuable species for investigating allelic variations with potentially important effects during evolution. Hybridizations involving various *R. rugosa* cultivars have generated offspring with desirable traits superior to those of either parent; this heterosis has been exploited in rose breeding programs. The current *R. rugosa* reference genome (454.78 Mb), which was assembled using PacBio circular consensus sequencing (CCS) data, is from a self-compatible and fruit-bearing variety (hereafter referred to as *R.rug*)[11]. In this study, we assembled a chromosome-

scale genome of a traditional Chinese *R. rugosa* variety, Hanxiang (HX), with two haplotypes fully represented. We also resequenced several leading *Rosa* accessions and close relatives to explore the genetic diversity among geographically distinct *Rosa* populations. Furthermore, we performed a weighted gene co-expression network analysis (WGCNA) and analyzed the haplotypic variations and allelic imbalances of *R. rugosa* varieties HX and 'Guomeigui' (GM), which have differing fragrance types, to characterize genes in the aroma synthesis pathway and examine allelic expression patterns. Our results provide insights into the mechanism underlying heterosis and the evolutionary history of *Rosa* species as well as the formation of rose floral scents.

Results

Genome assembly and annotation.

The HX (Fig. 1a) genome size was estimated to be approximately 456.19 Mb, with a heterozygosity level of 1.52% according to the 17-mer analysis (Supplemental Fig. 1a, Supplementary Table S1), which was consistent with the genome size revealed by the flow cytometry analysis (approximately 420 Mb) (Supplemental Fig. 1b, Supplementary Table S2). On the basis of 25.93 Gb PacBio HiFi data and 47.84 Gb Hi-C data (Table 1), the hifiasm software generated three assemblies, primary contigs (monoploid assembly), and fully phased contigs of two haplomes (hap1 and hap2). The three assemblies, with a total length that was close to the estimated genome size, were clustered into seven pseudochromosomes using high-throughput chromatin conformation capture (Hi-C) data (125× coverage) (Supplementary Table S3, Fig. 1b, Supplementary Table S4, Supplemental Fig. 2). The HX monoploid assembly consisted of 468.88 Mb, whereas HX hap1 included 453.63 Mb and HX hap2 comprised 440.53 Mb. The monoploid, hap1, and hap2 assemblies were 97.4%, 98.5%, and 98.6% complete according to the Benchmarking Universal Single-Copy Orthologs (BUSCO v.3) analysis (Supplemental Fig. 3). The LTR Assembly Index (LAI) values of the three assemblies exceeded 20 (i.e., reference level) (Supplemental Fig. 4). These results reflect the high accuracy and completeness of our *R. rugosa* genome assembly (Supplemental Fig. 5).

By integrating *de novo*, homologous (*Rosa chinensis*, *Rosa multiflora*, *R. rugosa*, *Fragaria vesca*, *Malus domestica*, and *Prunus persica*), and transcript (roots, stems, leaves, petals, stamens, pistils, sepals, and petioles) predictions, 36,023, 36,141, and 35,781 genes were predicted in the HX monoploid, HX hap1, and HX hap2 assemblies, respectively (Supplementary Tables S5, S6, S7, Supplemental Fig. 6). For the HX monoploid assembly, 37,534 (96.0%) genes were functionally annotated 37,455 (96.5%) and 37,181 (96.2%) genes in HX hap1 and hap2 were annotated (Supplementary Table S8). Transposable elements accounted for 242.88 Mb (51.8%) of the HX monoploid assembly, 228.55 Mb (50.38%) of the HX hap1 assembly, and 226.30 Mb (51.37%) of the HX hap2 assembly (Supplementary Table S9). Non-coding RNA sequences were also annotated (Supplementary Tables S10, S11, S12).

The SyRI [12] program was used to detect structural variants (SVs) and evaluate the divergence between the two haplomes. The 11,628 syntenic regions detected after aligning the two haplotypes covered

517.01 Mb (68.83%) of the anchored sequence. The syntenic regions consisted of 143 inversions, 8,991 translocations, 2,583 insertions, and 2,835 deletions, with 33 large SVs (> 100 kb) (Supplementary Table S13). Overall, the intragenomic diversity was estimated to be approximately 2.97%, which was greater than the diversity among out-crossing maize lines. The substantial variation was consistent with the high genomic heterozygosity and reflected the considerable diversity between the two haplomes.

We compared the syntenic patterns of the HX heterozygous genome with those of the *R.rug* and *R. chinensis* (*R.chi*) genomes (Fig. 1d). We determined that HX had highly conserved syntenic relationships with *R.rug* and *R.chi*, with closely matched chromosomes. Large SVs were detected among the three genomes (Fig. 1d; green blocks). Chromosomes 1, 3, and 7 of the HX heterozygous genome were similar to *R.chi* chromosomes 4, 1, and 7, respectively. However, compared with chromosome 1 of *R.rug*, the head of chromosome 1 of the HX heterozygous genome had large inversions. Moreover, two large SVs were detected between chromosome 7 of the HX heterozygous genome and *R.rug* chromosome 7.

Table 1
Details regarding the *R. rugosa* HX genome assembly and annotation

Assembly characteristics	Hap1	Hap2	Monoploid
Total length of contigs	453.62M	440.52M	468.87M
N50 length of contigs	7.98M	11.01M	24.41M
Total number of contigs	291	177	183
Longest contigs	16.45M	25.74M	54.40M
Total length of scaffolds	453.63M	440.53M	468.88M
N50 length of scaffolds	60.56M	58.78M	59.40M
Total number of scaffolds	172	99	154
Longest scaffolds	80.14M	76.04M	75.35M
Total gap size	11.90K	7.80K	2.90K
Total sequences anchored to the pseudochromosomes	443.16M	434.14M	431.47M
Number of annotated high-confidence genes	36,141	35,781	36,023
Percentage of transposon element sequences	50.38%	51.37%	51.80%
Complete BUSCOs	98.50%	98.60%	97.40%
Fragmented BUSCOs	0.70%	0.60%	0.70%
Missed BUSCOs	0.80%	0.80%	1.90%
LAI	21.91	21.4	21.13

Haplotype variations and allelic imbalance.

On the basis of the phased haplotypes, we separated 21,033 gene pairs (56.37%) with two defined alleles. The coding sequences of most allelic genes were highly similar (median = 0.99; Fig. 2a) and 81.92% of the allelic genes underwent purifying selection (Ka/Ks ratio < 1) (Fig. 2b). Additionally, 83.96% of the allelic genes contained at least one non-synonymous substitution that may have influenced protein functions (Fig. 2c).

Petal transcriptome data for HX were generated at four flower development stages (bud stage, initial opening period, bloom period, and fading period) to explore imbalanced allele expression. Allele-specific expression (ASE) was observed for 11.46–15.37% of the genes (false discovery rate < 0.05; Fig. 2d, Supplementary Table S14). Unstable ASE between the two haplomes was detected for 2,527, 1,730, 704, and 17 genes at one, two, three, and four stages, respectively. The expression levels of 372 genes were higher in hap1 than in hap2 at all stages, whereas 383 genes were more highly expressed in hap2 than in hap1 at all stages. The consistent and inconsistent allelic expression patterns suggested that allele functions may vary among flower development stages because of different regulatory pathways. We compared the hap1 and hap2 gene expression levels in stage 3 (i.e., most important flower development stage) at the chromosome level using sliding windows (Fig. 2e). Compared with the other genomic regions, genes that were differentially expressed between haplotypes were more common in the SV regions or flanking regions (5 kb upstream and downstream, $P = 0.033$), implying that SVs affected ASE. There were 21 significantly enriched biological pathways among the alleles specifically expressed in HX (stage 3), including glycerophospholipid metabolism, terpenoid backbone biosynthesis, ascorbate and aldarate metabolism, and inositol phosphate metabolism, which is related to 2-phenylethanol (2-PE) and citronellol biosynthesis (Supplementary Table S15, S16), suggesting that ASE may affect floral aroma substances in petals.

Rose population structure and pre-breeding improvement of the flower scent.

A total of 133 *Rosa* samples were selected for genome resequencing (average depth of 13×), including 28 *R. rugosa* accessions, 19 fragrant *Rosa* species, and 86 scented *R. hybrida* cultivars. We detected 13,741,097 high-quality SNPs in the 133 *Rosa* samples. The neighbor-joining phylogenetic tree constructed for the 133 *Rosa* samples and five peach accessions (i.e., outgroup) separated the samples into two groups, with group I including most of the *R. rugosa* accessions and group II including all of the scented *R. hybrida* cultivars and fragrant *Rosa* species (Fig. 3a). The neighbor-joining tree was constructed and the model-based clustering analysis (K = 2, 3, 4, 5, and 6) was completed on the basis of 1,195,460 filtered SNPs (Fig. 3a). Group II was further divided into three subclades, consistent with their geographical distribution patterns. Group II-1 mainly contained accessions of southwestern wild species (RSWs) from China, group II-2 mainly comprised scented *R. hybrida* cultivars (RHs), and group II-3 mainly included western varieties (RSs), primarily from Europe (e.g., *R. centifolia* and *R. damascena*) (Supplementary Table S17). These classifications were further supported by the principal component analysis (PCA) (Supplemental Fig. 7). The *R. rugosa* (RR) accessions, including wild species, landraces,

and improved cultivars (Fig. 3a), formed a monophyletic lineage, indicating that all currently grown *R. rugosa* accessions originated from a single domestication event (Fig. 3a). In addition, admixed ancestry was detected for some *Rosa* accessions, suggestive of introgression or gene flow during breeding.

The SMC + + method was used to reveal the fluctuations in the RR and RH population sizes from 0.7 Ma to 0.03 Ma (Fig. 3b). For both RR and RH, the effective population size decreased sharply during the Wangkun (0.7–0.5 Ma), Zhonglianggan (0.52–0.46 Ma), and final glaciation period (0.07–0.01 Ma), although there was a significant increase during the MIS 3c warm period (0.04–0.02 Ma) and then a stabilization with a slight increase in the RR effective population size. Accordingly, the trends in the RR and RH effective population sizes were roughly consistent with the climate changes during the Quaternary glacial period in China (Fig. 3b). A pairwise sequentially Markovian coalescent (PSMC) analysis showed that RRs had an earlier expansion as well as a lager effective population size than RHs (Fig. 3c).

Rosa rugosa is one of the original parents of modern cultivated rose species (i.e., RH)[13]. The nucleotide diversity (π) of the 133 accessions was 7.83×10^{-4}. The π was slightly larger for RH (8.91×10^{-4}) than for RR (5.06×10^{-4}), indicative of a lack of a domestication bottleneck (Fig. 3d), which was consistent with the results of the linkage disequilibrium (LD) analyses of each group (Fig. 3e). The relatively low LD of RH suggested that the dense coverage of the genome with many markers is ideal for a high-resolution population genetics analysis and that aroma traits were weakened during artificial breeding. Similar genetic diversity patterns were observed in other perennial crops, including grape (*Vitis vinifera*)[14] and apple[15], but not in peach[16,17].

In terms of the flower scent characteristics, 16 major components of the rose scent were detected by a targeted metabolite analysis, with several compounds contributing substantially to the scent, including 2-PE, citronellol, nerol, linalool, and rose ether. Compared with the other accessions, the RR petals contained more aroma substances, especially citronellol and rose ether, but their 2-PE content was less than that of RS petals (Fig. 3f, Supplementary Table S18). Therefore, RR flowers may be enhanced by modulating the 2-PE content.

Identification of conserved and/or divergent gene co-expression modules.

Both HX and GM were selected for the WGCNA to characterize genes in the aroma synthesis pathway. Notably, the 2-PE content was significantly higher in HX than in GM (Supplementary Table S20). To investigate the HX and GM gene regulatory networks during flower development stages (bud stage, initial opening period, bloom period, and fading period; a total of 24 samples) (Fig. 4a), we identified co-expressed gene sets via a WGCNA, with the 2-PE content selected as the phenotype. The gene regulatory network analysis revealed several major subnetworks involving interactions among genes with similar expression profiles (i.e., co-expression modules). Nineteen modules (9–6,839 genes) were identified in HX and GM (Fig. 4b, c, Supplementary Table S22). The module eigenvalue (e) was used to determine gene expression patterns in each sample. A heat map of the expression patterns revealed modules

significantly related to specific samples. Interestingly, 10 co-expression modules in HX and 10 co-expression modules in GM were highly correlated ($e \geq 0.30$) with flower development stages (Fig. 4c, Supplementary Table S21). The genes in the tan and steelblue modules were more highly expressed in the HX petals than in the GM petals (Fig. 4d).

The petal transcriptional changes during flower development stages were compared between HX and GM to clarify the relationships between the 2-PE synthesis-related genes. The production of 2-PE is directly related to the phenylalanine, tyrosine, and tryptophan biosynthetic pathway as well as the phenylalanine metabolic pathway. The expression levels of 34 genes encoding 16 enzymes changed in HX and GM at different flower development periods. In this pathway, the expression levels of the primary amine oxidase (PAO)-encoding genes *RrHX1G119800* and *RrHX1G204700* and the L-tryptophan decarboxylase (L-TDC)-encoding gene *RrHX2G284700* were up-regulated in the third and fourth flowering stages, which was in contrast to the down-regulated expression of the genes encoding other pathway enzymes (Fig. 4e).

Variations in three ASE genes related to PE synthesis in the population and in other genomes.

Among above 34 genes encoding 16 enzymes in 2-PE synthesis pathway, we found the PAO-encoding genes *RrHX1G204700* and *RrHX1G119800* and the L-TDC-encoding gene *RrHX2G284700* were three ASE genes significantly associated with the PE content (Fig. 4e, Supplementary Table S23) and some non-synonymous SNPs in these genes resulted in two major haplotypes (Fig. 5). There were two non-synonymous SNPs in *RrHX1G204700* and the frequency of ‘GA’ increased in samples with increased PE contents (Fig. 5b). The accessions with ‘GA’ had a significantly higher PE content (11.4039 μg/g, $P < 0.01$) than those with ‘AC’ (3.2556 μg/g) (Supplementary Table S24). Thus, ‘GA’ was the favorable haplotype. Moreover, the one non-synonymous SNP in *RrHX2G284700* and *RrHX2G119800* significantly associated with the PE content resulted in two major haplotypes (Fig. 5a, c). In *RrHX2G284700*, ‘A’ was the favorable haplotype. (Fig. 5c). The average 2-PE content in accessions with ‘A’ (17.6288 μg/g) was significantly higher than that in accessions with ‘T’ (6.291 μg/g) ($P < 0.05$) (Supplementary Table S25).. In *RrHX2G119800* (Fig. 5a), ‘T’ was the favorable haplotype and the average 2-PE content in the accessions with ‘T’ (12.3514 μg/g) was significantly higher than that in the accessions with ‘A’ (5.7462 μg/g) ($P < 0.05$) (Supplementary Table S26).

The expression patterns of the genes encoding PAO and L-TDC differed between hap1 and hap2. The genomes of these two haplotypes were used as the reference material for investigating the asymmetrical expression of the screened alleles (Fig. 5). Three key genes, namely *PAO* (*RrHX7G119800* and *RrHX1G204700*) and *L-TDC* (*RrHX2G284700*), had different allelic expression patterns in different periods, suggesting the functions of these alleles may vary depending on the developmental stage. In HX and GM, which differ in terms of their fragrance and PE content, there were more inconsistent ASE genes than consistent ASE genes. Moreover, compared with HX, GM had a lower PE content but more consistent ASE genes (Supplementary Table S14). We speculated that inconsistent ASE genes influence 2-PE synthesis in *R. rugosa*.

A total of 110 SNPs, 24 insertions, and 12 deletions in the PAO-encoding gene *RrHX1G204700* (gene body and promoter region) were detected following the comparison between HX and *R.chi*, but the comparison between HX and *R.rug* identified only one SNP and one deletion (Fig. 5). Additionally, more genetic structural variation in the PAO-encoding gene *RrHX7G119800* and L-TDC-encoding gene *RrHX1G284700* were detected by the comparison between HX and *R.chi*, than between HX and *R.rug*. Thus, the three genes were more similar between HX and *R.rug* than between HX and *R.chi* (Supplemental Fig. 8). The variations in these ASE genes may help to explain the diversity in rose flower fragrances.

Discussion

The cultivation of rose species in China started in the Western Han Dynasty (202 BC–9 AD). Clonal propagation for approximately 2,000 years resulted in the accumulation of substantial somatic mutations in the genome, allowing us to separate the two haplotypes and identify allelic imbalances. The ASE genes were classified as consistent ASE genes and inconsistent ASE genes (i.e., direction-shifting pattern). The asymmetrical expression of selected alleles was studied on the basis of the HX and GM transcriptome data at four flowering stages. In contrast to GM (with a low PE content), there were considerably more inconsistent ASE genes than consistent ASE genes in HX (10,596 *vs* 1,247), suggesting that the overdominance effect is a major reason for the high heterozygosity of the *R. rugosa* genome. The large abundance of inconsistent ASE genes in rose plants is likely caused by the accumulation of somatic mutations over a long period of clonal propagation, similar to what occurred in hybrid rice[18], but not in tea plants[19]. In addition, the haplotype frequency changes of three key ASE genes (encoding *PAO* and *L-TDC*) related to PE synthesis in different groups and the differences in these three genes between the HX genome and other rose genomes were explored. The results suggested that ASE frequency changes affected gene expression, which resulted in phenotypic alterations.

The genus *Rosa* comprises approximately 200 species, more than half of which are polyploids. Rose species have been modified by extensive reticulate evolution and interspecific hybridization, introgression, and polyploidization[20]. Approximately 8–20 *Rosa* species contributed to the genetic make-up of modern rose cultivars, including *R. rugosa*, *R. chinensis*, *R. multiflora*, and *R. gigantea* as well as European rose species *R. moschata*, *R. gallica*, *R. canina*, and *R. phoenicia*[21,22]. Our results revealed gene introgressions among *R. rugosa*, *R. hybrida*, and other *Rosa* species, similar to the introgression in walnut populations[23]. There were more introgressions in *R. hybrida* and other *Rosa* populations than in *R. rugosa* populations. This may be related to the affinity among the three groups and the selective crossbreeding by humans[2]. Climate changes during the Quaternary glacial and interglacial periods directly influenced the distribution of plant taxa and the size of plant habitats[24]. The RR and RH effective population sizes tended to decrease sharply during the glacial period and stabilized or increased during the interglacial period, which is in accordance with the related findings for orchids[25], *Rhododendron*[26,27], and other plant groups.

The major scent components of *R. rugosa* and European rose flowers are 2-PE and several monoterpenes (e.g., rose oxide, geraniol, and nerol). In contrast, the *R. chinensis* floral fragrance is primarily due to lipid-

derived alcohols and esters (e.g., hexenol and hexenyl acetate) and aromatic compounds (e.g., 3,5-dimethoxytoluene and 1,3,5-trimethoxybenzene)[28]. Previous research confirmed 2-PE belongs to the second largest class of floral compounds (i.e., phenylpropanes) that produce a “honey-like” odor[28]. Several natural 2-PE biosynthesis-related pathways have been identified, including the Ehrlich pathway, phenylethylamine pathway, and phenylpyruvate pathway. We also determined that *RrHX7G119800* (PAO), *RrHX1G204700* (PAO), and *RrHX2G284700* (L-TDC) encode phenylethylamine pathway enzymes that promote 2-PE production at different stages. Future studies should aim to identify universal candidate genes with similar functions in different species, such as tobacco, grass, and petunia, while also clarifying the regulation of the core genes affecting phenylethanol synthesis in rose species, which may be relevant for optimizing the 2-PE content in rose flowers.

In conclusion, this study provides important insights into rose genome evolution, allelic imbalances, and population genetics, which may form the basis of future research conducted to enhance rose floral aromas through breeding. Our newly developed genomic resources may be applicable for molecular biology research and may help to shorten the breeding cycle by enabling breeders to target specific genes in *R. rugosa* and other aromatic crops.

Methods

Plant materials

The sequenced rose (*Rosa*) accessions used in this study were grown and collected from May to July 2021. Specifically, the materials were cultivated in the *Rosa* germplasm resource nursery of the Beijing Academy of Agriculture and Forestry Sciences at Beilangzhong, Zhaoquanying, and Shunyi in Beijing, China.

Genome Sequencing

For the *de novo* assembly, samples were collected from a single plant. For the CCS analysis, genomic DNA was extracted from *in vitro* seedlings using the DNeasy Plant Mini Kit (Qiagen, Shanghai, China). A 15-kb library was constructed and sequenced using the Pacific Bioscience Sequel II platform (Annoroad Gene Technology). A total of 16.61 Gb CCS reads with an N50 value of 18.68 kb were generated using the ccs software (v.3.0.0) (https://github.com/pacificbiosciences/unanimity/; –min-passes 3 –min-length 10000 –max-length 1000000 –min-rq 0.99). For the Hi-C analysis, leaf tissues were fixed in 1% formaldehyde before constructing libraries. After verifying the quality of the libraries was acceptable, they were sequenced using the Illumina HiSeq NovaSeq platform as previously described[29]. Briefly, the DpnII restriction enzyme was used for digesting chromatin. Libraries constructed using the Illumina TruSeq DNA Sample Prep Kit were sequenced using the Illumina NovaSeq Xten system to produce 2 × 150-bp reads (25.93 Gb). The RNA-seq analysis was completed using root, stem, leaf, petal, stamen, pistil, sepal, and petiole samples collected for each accession. Total RNA was isolated using the RNAprep Pure Plant

Kit (TIANGEN). The RNA-seq libraries were constructed according to a published method[30] and then sequenced on the NovaSeq platform.

Genome Size Estimation

The genome size was estimated on the basis of a k-mer frequency analysis. The k-mer distribution depends on genome characteristics and follows Poisson's distribution. Before the genome was assembled, the 17-mer distribution of WGS reads was determined using Jellyfish (v.2.2.6)[31]. Additionally, GCE[32] was used to estimate genome characteristics.

Genome Assembly

The contig genome was assembled using HiFi reads and Hi-C reads in Hi-C models. The following parameters of the hifiasm software were applied: '-l 2 -k 51 -w 51 –write-paf –write-ec –h1 –h2'. To comprehensively characterize the genome, we used p_ctg.fa as the monoploid contig result and hic.hap1.p_ctg.fa and hic.hap2.p_ctg.fa as the hap 1 and hap 2 contig results. The contigs were anchored to chromosomes using the Hi-C reads, which were aligned to the contigs using HICUP (v.0.7.3)[33], yielding an alignment BAM file. The contigs were clustered using the ALLHiC[34] algorithm (.ALLHiC_partition -e GATC -k 7). Finally, the assembled genomes were manually corrected using Juicebox (v.1.11.08)[35]. Furthermore, BUSCO[36] and LAI[37] were used to determine the completeness on the basis of the Embryophyta Plant database and the full-length long terminal repeat retrotransposons, respectively.

Genome Annotation

Repeating elements were predicted via *de novo* and homology-based approaches. For the homology-based approach, the RepBase database was screened for sequence matches (http://www.girinst.org/repbase), after which repeating elements were predicted using RepeatProteinMask (http://www.repeatmasker.org/). For the *de novo* annotation, LTR_FINDER[38], RepeatScout (http://www.repeatmasker.org/), and Repeat-Modeler (http://www.repeatmasker.org/RepeatModeler.html) were used to construct a *de novo* library, which was then annotated using RepeatMasker (http://repeatmasker.org/).

A comprehensive strategy combining *ab initio* predictions, protein-based homology searches, and RNA sequencing was used to elucidate gene structures. Protein sequences from *R. chinensis*, *R. multiflora*, *R. rugosa*, *F. vesca*, *M. domestica*, and *P. persica* were aligned to the corresponding genome using WUblast[39], with an E-value cutoff of 1e-5. The resulting hits were conjoined using the Solar software[40]. GeneWise[41] was used to predict the precise gene structures of the corresponding genomic regions for each WUblast hit. Gene structures created by GeneWise were denoted as Homo-set (homology-based prediction gene set). Gene models created using PASA[42] were denoted as the PASA-T-set and used as the

training data for the following five *ab initio* gene prediction programs that were used to predict coding regions in the repeat-masked genome: Augustus (v.2.5.5)[43], Genscan (v.1.0)[44], GeneID[45], GlimmerHMM (v.3.0.1)[46], and SNAP[47]. The RNA-seq data (Supplementary Table S27) were mapped to the assembly using TopHat (v.2.0.8)[48] and then Cufflinks (v.2.1.1)[49] was used to assemble the transcripts into gene models (Cufflinks-set). In addition, the gene models predicted from the Trinity-assembled transcripts using PASA were designated as the PASA-T-set (PASA Trinity set). Gene model evidence from the Homo-set, PASA-Iso-set, Cufflinks-set, PASA-T-set, and *ab initio* data was combined using EVidenceModeler[50] into a non-redundant set of gene annotations. The weight of each type of evidence was as follows: PASA-T-set > Homo-set > Cufflinks-set > Augustus > GeneID = SNAP = GlimmerHMM = Genscan.

The predicted protein sequences were functionally annotated using the following five databases: NR, InterPro, GO, KEGG, and Swiss-Prot. The InterPro and GO databases were searched using the following parameters of InterProScan[51]: "-f TSV -dp -gotermes -iprlookup". The NR, Swiss-Prot, and KEGG databases were searched using BLAST, with an E-value cut-off of 1×10^{-5}. The results from these database analyses were concatenated.

Sv Identification

Haplotype 1 was aligned to hap 2 using the following parameters of mummer: –maxgap = 500 –mincluster = 500. The raw alignments were filtered using the following parameters of delta-filter: -i 95 -l 1000 -1. After using the "show-coords" command, the filtered results were converted to an easy-to-read file. This file was used to detect structural variations using the following default parameters of the SyRI pipeline: –no-chrmatch –nosnp.

Identification Of Alleles

Interhaplotype syntenic blocks were identified using the default parameters of MCScan. The paired genes corresponding to the same blocks were considered to be alleles.

Dominant Allele Expression

Tissues (e.g., leaves and flowers) in different periods were collected from mature plants for an RNA-seq analysis. The generated RNA-seq reads were trimmed using Fastp and then mapped to the combined genomes (hap1 and hap2) using HISAT2. The FPKM values were calculated on the basis of unique mapped reads using the featureCounts software and a homemade R script. To analyze dominant allele expression, we extracted 21,033 genes with known alleles. The allele pairs were compared to analyze differential expression.

Snp And Small Indel Calling

We collected Illumina resequencing data for 133 newly sequenced *Rosa* accessions (average depth of 13×). The raw resequencing data were screened for high-quality reads using the default settings of fastp (v.0.20.1)[52]. To detect SNPs, Illumina short reads were aligned to the rose genome using BWA-MEM; PCR duplicates were removed using Picard (v.1.118) (http://broadinstitute.github.io/picard/). The SNPs and InDels were identified using HaplotypeCaller from the Genome Analysis Toolkit (v.4.1.5.0)[53] and then filtered as previously described[54]. The SNPs with a read depth < 5 and non-biallelic SNPs were eliminated.

Linkage Disequilibrium

To estimate and compare the LD decay patterns in each population, we calculated the mean squared correlation coefficient (r^2) values for any two SNPs within 500 kb using PopLDdecay (v.3.41)[55]. A 500-bp bin was used to generate the plot.

SMC++

The SMC + + [56] (v.1.13) program was used to estimate the divergence time and historical Ne among different rose groups. Generations were calculated using a mutation rate of 9.48×10^{-9} per synonymous site for each generation, with a generation time of 2 years.

Preparation Of Transcriptome Sequencing Samples

Petals were collected from HX and GM plants at four flowering stages for the subsequent RNA extraction. Analyses were completed using the following replicates: HX 1–1, HX1-2, HX1-3, HX2-1, HX2-2, HX2-3, HX3-1, HX3-2, HX3-3, HX4-1, HX4-2, HX4-3, GM1-1, GM1-2, GM1-3, GM2-1, GM2-2, GM2-3, GM3-1, GM3-2, GM3-3, GM4-1, GM4-2, and GM4-3.

Rna Extraction, Cdna Library Construction, And Illumina Sequencing

Total RNA was extracted from the leaf samples and used to prepare sequencing libraries for the RNA-seq analysis. Specifically, total RNA was isolated using the TRIzol reagent according to the manufacturer's protocol (TaKaRa, Dalian, China). Eukaryotic mRNA in the extracted total RNA was enriched using oligo-(dT) beads. The cDNA was then purified using the QIAquick PCR Purification Kit. A poly-(A) sequence was added to the blunt-ended cDNA, which was then ligated to the Illumina sequencing linker. The ligation product of the desired size was selected by agarose gel electrophoresis, amplified by PCR, and sequenced using the Illumina HiSeq™ 4000 system at Gene Denovo Biotechnology Co. (Guangzhou, China).

Read Mapping And Differential Gene Expression Analysis

To obtain high-quality clean reads for the subsequent sequence assembly and analysis, the raw reads were filtered by removing reads with adapters, reads containing more than 10% unknown nucleotides, and reads with more than 40% low-quality bases (Q-value ≤ 10). The retained high-quality clean reads were mapped to rRNA to identify residual RNA reads. After eliminating the rRNA reads, the remaining high-quality clean reads were mapped to the reference transcriptome using the default parameters of Bowtie2[57]. The mapping rate was calculated using the following equation: mapping rate = (number of unique mapped reads + number of multiple mapped reads)/total number of reads.

Gene expression levels were calculated and normalized as RPKM (reads per kilobase per million reads) values[58] using the following formula: RPKM = 106 × C/(NL/103), where C is the number of reads uniquely aligned to a specific gene, N is the total number of reads uniquely aligned to all genes, and L is the number of bases in the specific gene. Calculating gene expression according to the RPKM value eliminated the influence of gene length and sequencing data abundance. Therefore, the calculated gene expression levels could be used for a direct comparison of gene expression among samples.

The reliability of the results was assessed by a correlation analysis of two parallel experiments. The reproducibility between samples was determined on the basis of the correlation coefficient (i.e., a correlation coefficient close to 1 reflects the high reproducibility between two parallel experiments). The R package models (http://www.r-project.org/) were used to perform a PCA, in which hundreds of thousands of correlated variables (gene expression) were converted into a set of values of linearly uncorrelated variables called principal components. The PCA is largely used to reveal the structure/relationship among samples/data.

The edge R package (http://www.r-project.org/) was used to identify differentially expressed genes (DEGs) across samples or groups. An expression level fold-change ≥ 2 and a false discovery rate < 0.05 were used as the thresholds for identifying significant DEGs, which were then subjected to GO and KEGG pathway enrichment analyses.

Gene Expression Analysis

Gene expression patterns were analyzed to cluster the genes similarly expressed in multiple samples (i.e., at least three at a specific time-point). To examine the DEG expression patterns, the expression data for each sample were normalized to 0, $\log_2$(v1/v0), and $\log_2$(v2/v0) and then clustered using the Short Time-series Expression Miner software[59]. The parameters were set as follows: maximum unit change in model profiles between time-points was 1; maximum output profiles was 20 (similar profiles were merged); and the minimum DEG expression level fold-change was ≥ 2.0. A *P*-value ≤ 0.05 was used as the threshold for identifying significant cluster profiles. The DEGs in the profiles were subjected to GO and KEGG pathway enrichment analyses. An adjusted *P*-value based on a false discovery rate ≤ 0.05[60] was used to identify the significantly enriched GO terms and KEGG pathways.

Gas Chromatography–mass Spectrometry And Statistical Analyses Of Targeted Metabolomics And Volatile Metabolomics Data

Fresh petal samples were collected from 27 *R. rugosa* cultivars, 43 scented *R. hybrida* cultivars, and seven fragrant *Rosa* species and weighed, frozen in liquid nitrogen, and stored at − 80°C for the targeted metabolomics analysis. The samples were ground to a powder in liquid nitrogen and transferred to a 20 mL headspace vial (Agilent, Palo Alto, CA, USA) containing NaCl saturated solution to inhibit enzymatic reactions. The vials were sealed using crimp-top caps with TFE–silicone headspace septa (Agilent) prior to the HS-SPME. Samples were incubated at 60°C for 10 min. A 65-μm divinylbenzene/carboxen/polydimethylsiloxane fiber (Supelco, Bellefonte, PA, USA) was then inserted into the headspace of each vial for 20 min at 60°C. The VOCs were desorbed from the fiber coating in the injection port of the GC apparatus (Model 8890; Agilent) at 250°C for 5 min in the splitless mode. The VOCs were identified and quantified using the Model 8890 GC system (Agilent) and the 5977B MS system (Agilent) equipped with a 30 m × 0.25 mm × 0.25 μm DB-5MS (5% phenyl-polymethyl siloxane) capillary column. We used helium as the carrier gas at a linear velocity of 1.0 mL/min. The injector temperature was kept at 250°C and the detector temperature was set at 280°C. The oven temperature was maintained at 40°C for 5 min and then increased to 280°C at 5°C/min and held for 5 min. We recorded the mass spectra in the electron impact (EI) ionization mode at 70 eV. The quadrupole mass detector, ion source, and transfer line temperatures were set at 150, 230, and 280°C, respectively. Mass spectra were scanned in the range m/z 30–350 amu at 1 s intervals. Volatile compounds were identified by comparing the mass spectra with the data system library (MWGC) and linear retention indices.

We extracted citronellol, phenethyl alcohol, farnesol, nerol, and rose oxide from petals by HS-SPME for the targeted metabolomics analysis using the GC-MS system[61,62]. The analysis was completed as previously described[2,63–65]. The metabolites were sorted according to their content and concentration. On the basis of the International Standard for Oil of Rose (ISO 9843: 2003), we compared the most important metabolites with published data for the following 16 target metabolites: eight terpenes (α-farnesene, nerolidol, nerol, farnesol, rose ether, citronellal, citronellol, and linalool), four esters (neryl acetate, citronellyl acetate, phenethyl acetate, and geranyl acetate), two phenols (eugenol and methyleugenol), one alcohol (phenethyl alcohol), and one heterocyclic compound (*trans*-linalool oxide). We prepared standard curves for the 16 target substances and quantified their contents.

The HX and GM *R. rugosa* accessions with differing fragrances were selected for the volatile metabolomics analysis at four flower development stages. We used the Mass Hunter software (Agilent) to extract and analyze the original data for the volatile metabolites identified by GC-MS and then analyzed the data qualitatively, quantitatively, and statistically. We prepared the quality control (QC Mix) sample by combining the extracts of all samples to assess the repeatability of the results.

Co-expression Network Analysis For Constructing Modules

Co-expression networks were constructed using the WGCNA (v.1.47) package in R [66]. After filtering genes (i.e., removing more than 40% of the genes according to their RPKM values), the gene expression values were imported into WGCNA to construct co-expression modules using the automatic network construction function blockwiseModules. The default settings were used, with the following exceptions: power was 10, TOM Type was unsigned, mergeCutHeight was 0.7, and minModuleSize was 50. Genes were clustered into 21 correlated modules. For each module, a summary profile (module eigengene) was calculated via a PCA. To evaluate the association between the modules and stage-specific expression in each cultivar, we determined the correlation between each module eigengene and the 2-PE content in petals as previously described (Downs et al., 2013)[67].

Genotyping Of The Population Comprising Samples With Differing 2-pe Contents

The ASE genes related to phenylethanol synthesis were screened. The SNPs resulting in non-synonymous mutations were extracted, and the SNPs with a heterozygosity greater than 0.2 and a maf less than 0.05 were filtered. The filtered SNPs were used for the genotyping of the samples with varying 2-PE contents. The genes with different major haplotype frequencies in the two populations were selected. The corresponding phenotypes of all samples with different haplotypes were examined using the *t*-test, after which the genes associated with significantly different phenotypes were retained.

Declarations

Author Contributions

Xi Cheng: Conceptualization, Methodology, Formal analysis, Resources, Writing and Editing, Funding acquisition. Dan Gao: Investigation, Data Curation, Writing and Editing. Hongli Wang: Formal analysis. Guoliang Wang: Data Curation. Dongliang Chen: Writing- Reviewing, Validation. Chang Luo: Writing- Reviewing. Hua Liu: Writing- Reviewing. Tianyi Wang: Investigation, Data Curation,. Chengzhi Jiao: Supervision. Kezhong Zhang: Writing- Reviewing. Beibei Jiang: Writing- Reviewing, Funding acquisition. Conglin Huang: Supervision, Funding acquisition, Project administration.

Acknowledgment

This work was supported by funds from the Beijing Academy of Agriculture and Forestry Sciences (PT2023-03) and the Beijing Innovation Consortium of Agriculture Research System (BAIC09-2023). We thank Professor Zhangjun Fei (Boyce Thompson Institute, Cornell University) for critical reading of this manuscript and Xuepeng Sun Zhejiang A&F University for his suggestions on data analysis. We thank Liwen Bianji (Edanz) (www.liwenbianji.cn) for editing the English text of a draft of this manuscript.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Raguso, R.A. Wake up and smell the roses: the ecology and evolution of floral scent. Annual review of ecology, evolution, and systematics, 549–569 (2008).
2. Raymond, O. *et al.* The Rosa genome provides new insights into the domestication of modern roses. Nature genetics 50, 772–777 (2018).
3. Tholl, D. & Gershenzon, J. The flowering of a new scent pathway in rose. Science 349, 28–29 (2015).
4. Committee, N.P. Pharmacopoeia of the People's Republic of China. *Part* 1, 188–189 (2015).
5. Orhan, G., Orhan, I., Subutay-Oztekin, N., Ak, F. & Sener, B. Contemporary anticholinesterase pharmaceuticals of natural origin and their synthetic analogues for the treatment of Alzheimer's disease. Recent Patents on CNS Drug Discovery (Discontinued) 4, 43–51 (2009).
6. Hancianu, M., Cioanca, O., Mihasan, M. & Hritcu, L. Neuroprotective effects of inhaled lavender oil on scopolamine-induced dementia via anti-oxidative activities in rats. Phytomedicine 20, 446–452 (2013).
7. Senol, F.S. *et al.* A mechanistic investigation on anticholinesterase and antioxidant effects of rose (Rosa damascena Mill.). Food research international 53, 502–509 (2013).
8. Linck, V.M. *et al.* Effects of inhaled Linalool in anxiety, social interaction and aggressive behavior in mice. Phytomedicine 17, 679–683 (2010).
9. Wang, Y. *et al.* Enhanced antibacterial activity of eugenol-entrapped casein nanoparticles amended with lysozyme against gram-positive pathogens. Food Chemistry 360, 130036 (2021).
10. Gunaseelan, S. *et al.* The preventive effect of linalool on acute and chronic UVB-mediated skin carcinogenesis in Swiss albino mice. Photochemical & Photobiological Sciences 15, 851–860 (2016).
11. Chen, F. *et al.* A chromosome-level genome assembly of rugged rose (Rosa rugosa) provides insights into its evolution, ecology, and floral characteristics. Horticulture research 8(2021).
12. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome biology 20, 1–13 (2019).
13. Bendahmane, M., Dubois, A., Raymond, O. & Bris, M.L. Genetics and genomics of flower initiation and development in roses. Journal of experimental botany 64, 847–857 (2013).
14. Zhou, Y., Massonnet, M., Sanjak, J.S., Cantu, D. & Gaut, B.S. Evolutionary genomics of grape (Vitis vinifera ssp. vinifera) domestication. *Proceedings of the National Academy of Sciences* 114, 11715–11720 (2017).
15. Duan, N. *et al.* Genome re-sequencing reveals the history of apple and supports a two-stage model for fruit enlargement. Nature Communications 8, 1–11 (2017).

16. Li, Y. *et al.* Genomic analyses of an extensive collection of wild and cultivated accessions provide new insights into peach breeding history. Genome biology 20, 1–18 (2019).
17. Yu, Y. *et al.* Population-scale peach genome analyses unravel selection patterns and biochemical basis underlying fruit flavor. Nature communications 12, 1–13 (2021).
18. Shao, L. *et al.* Patterns of genome-wide allele-specific expression in hybrid rice and the implications on the genetic basis of heterosis. *Proceedings of the National Academy of Sciences* 116, 5653–5658 (2019).
19. Zhang, X. *et al.* Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant Camellia sinensis. Nature Genetics 53, 1250–1259 (2021).
20. Fougère-Danezan, M., Joly, S., Bruneau, A., Gao, X.-F. & Zhang, L.-B. Phylogeny and biogeography of wild roses with specific attention to polyploids. Annals of Botany 115, 275–291 (2015).
21. De Vries, D. & Dubois, L.A. Rose breeding: past, present, prospects. in *II International Rose Symposium 424* 241–248 (1995).
22. Reynders-Aloisi, S. & Bollereau, P. Characterisation of genetic diversity in genus Rosa by randomly amplified polymorphic DNA. in *II International Rose Symposium 424* 253–260 (1995).
23. Ding, Y.-M. *et al.* Population-genomic analyses reveal bottlenecks and asymmetric introgression from Persian into iron walnut during domestication. Genome biology 23, 1–18 (2022).
24. Petit, R.J. *et al.* Fagaceae trees as models to integrate ecology, evolution and genomics. New Phytologist 197, 369–371 (2013).
25. Zhang, W. *et al.* Genome sequence of Apostasia ramifera provides insights into the adaptive evolution in orchids. BMC genomics 22, 1–12 (2021).
26. Ma, H. *et al.* Chromosome-level genome assembly and population genetic analysis of a critically endangered rhododendron provide insights into its conservation. The Plant Journal 107, 1533–1545 (2021).
27. Wang, K. *et al.* The uplift of the Hengduan Mountains contributed to the speciation of three Rhododendron species. Global Ecology and Conservation 35, e02085 (2022).
28. Shi, S. & Zhang, Z. Genetic and biochemical aspects of floral scents in roses. International Journal of Molecular Sciences 23, 8014 (2022).
29. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science* 326, 289–293 (2009).
30. Liu, Z. *et al.* Comparison of genetic diversity between Chinese and American soybean (Glycine max (L.)) accessions revealed by high-density SNPs. Frontiers in plant science 8, 2014 (2017).
31. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27, 764–770 (2011).
32. Liu, B. *et al.* Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv preprint arXiv:1308.*2012 (2013).
33. Wingett, S. *et al.* HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research* 4(2015).

34. Zhang, J. *et al.* Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. Nature genetics 50, 1565–1573 (2018).
35. Burton, J.N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nature biotechnology 31, 1119–1125 (2013).
36. 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).
37. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic acids research 46, e126-e126 (2018).
38. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic acids research 35, W265-W268 (2007).
39. She, R., Chu, J.S.-C., Wang, K., Pei, J. & Chen, N. GenBlastA: enabling BLAST to identify homologous gene sequences. Genome research 19, 143–149 (2009).
40. Yu, X.-J., Zheng, H.-K., Wang, J., Wang, W. & Su, B. Detecting lineage-specific adaptive evolution of brain-expressed genes in human using rhesus macaque as outgroup. Genomics 88, 745–751 (2006).
41. Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. Genome research 14, 988–995 (2004).
42. Haas, B.J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic acids research 31, 5654–5666 (2003).
43. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic acids research 34, W435-W439 (2006).
44. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. Journal of molecular biology 268, 78–94 (1997).
45. Guigo, R. Assembling genes from predicted exons in linear time with dynamic programming. Journal of Computational Biology 5, 681–702 (1998).
46. Majoros, W.H., Pertea, M. & Salzberg, S.L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 20, 2878–2879 (2004).
47. Korf, I. Gene finding in novel genomes. BMC bioinformatics 5, 1–9 (2004).
48. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology 14, 1–13 (2013).
49. Ghosh, S. & Chan, C.-K.K. Analysis of RNA-Seq data using TopHat and Cufflinks. in *Plant bioinformatics* 339–361 (Springer, 2016).
50. Haas, B.J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome biology 9, 1–22 (2008).
51. Hunter, S. *et al.* InterPro: the integrative protein signature database. Nucleic acids research 37, D211-D215 (2009).
52. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884-i890 (2018).

53. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 20, 1297–1303 (2010).
54. Guo, S. *et al.* Resequencing of 414 cultivated and wild watermelon accessions identifies selection for fruit quality traits. Nature genetics 51, 1616–1623 (2019).
55. 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).
56. Terhorst, J., Kamm, J.A. & Song, Y.S. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nature genetics 49, 303–309 (2017).
57. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25, 1966–1967 (2009).
58. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods 5, 621–628 (2008).
59. Ernst, J. & Bar-Joseph, Z. STEM: a tool for the analysis of short time series gene expression data. BMC bioinformatics 7, 1–11 (2006).
60. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) 57, 289–300 (1995).
61. Feng, Y. *et al.* Gas chromatography-mass spectrometry analysis of floral fragrance-related compounds in scented rose (Rosa hybrida) varieties and a subsequent evaluation on the basis of the analytical hierarchy process. Plant Physiology and Biochemistry 185, 368–377 (2022).
62. Cheng, X. *et al.* Evaluation of Rosa germplasm resources and analysis of floral fragrance components in R. rugosa. Frontiers in plant science 13(2022).
63. Zhou, Y. *et al.* The analysis of aromatic compounds from Rosa chinensis ‘Pallida’, R. damascene, R. centifolia by GC/MS. (2015).
64. Yuan, N. *et al.* Determination of Aromatic Components of Rosa davurica Pall, by Headspace Solid Phase Microextraction Combined with GC-MS. Medicinal Plant 9(2018).
65. Yun, M.M., Bao-Yin, L.I. & Zhou, X.M. Determination of aromatic components of flower in Rosa rugosa Thunb. by the static headspace and gas chromatography-mass spectrometry technology. Science and Technology of Food Industry (2016).
66. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics 9, 1–13 (2008).
67. Downs, G.S. *et al.* A developmental transcriptional network for maize defines coexpression modules. Plant physiology 161, 1830–1843 (2013).

Figures

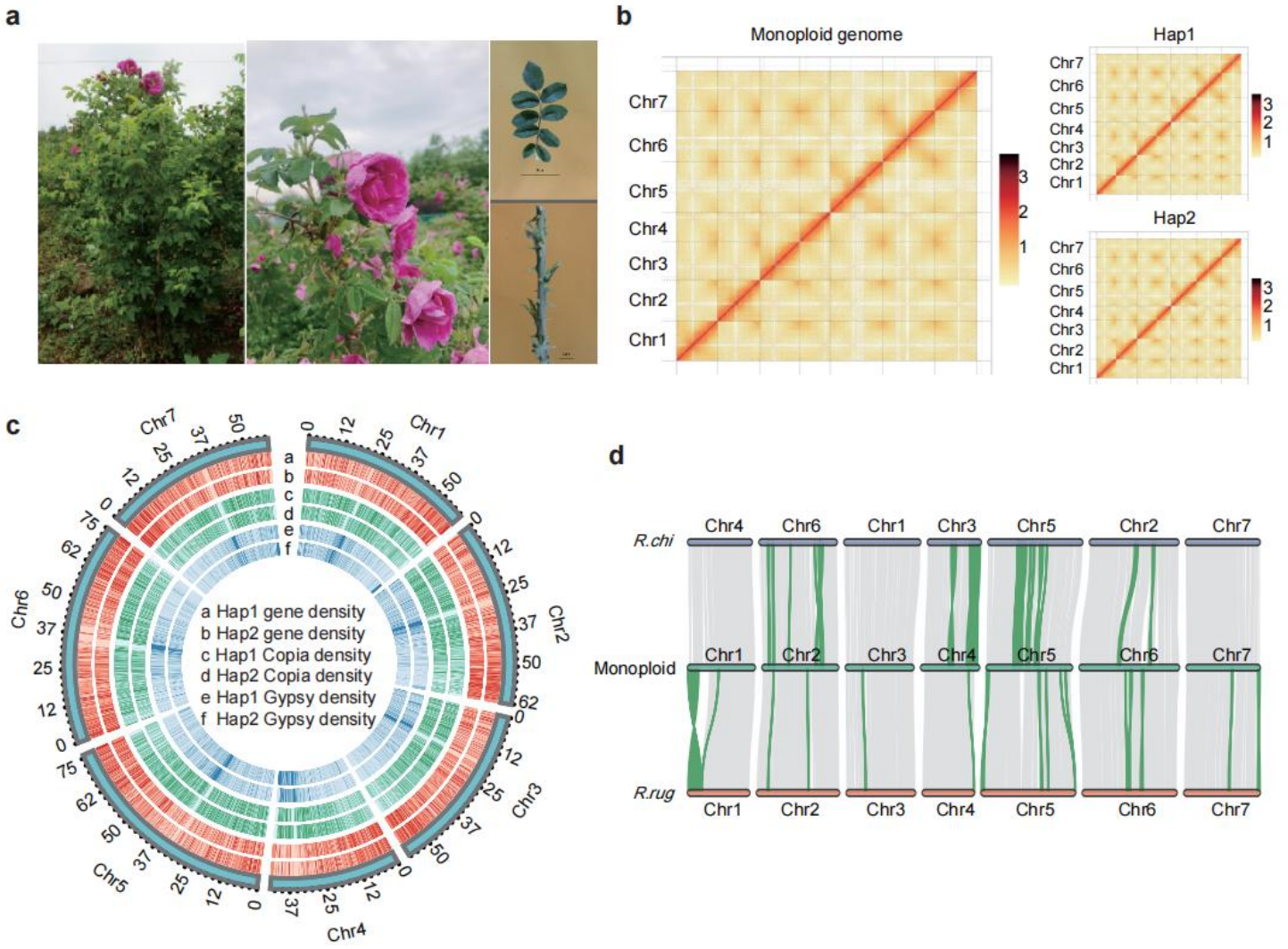

Figure 1

Assembly, composition, and evolution of the HX genome. a, Images of *R. rugosa* HX plant parts. Scale bar, 5 cm (upper panel) and 1 cm (lower panel). b, Contact map of the Hi-C links among seven pseudochromosomes. c, *R. rugosa* HX genomic features presented in a Circos diagram of hap1 and hap2. The circles (outer to inner) indicate gene density (a: hap1, b: hap2), Copia density (c: hap1, d: hap2), and Gypsy density (e: hap1, d: hap2). d, Analysis of the synteny between the *R.chi*genome assembly, monoploid (HX) genome assembly, and *R.rug* genome assembly. Green blocks represent large structural variations among the three genomes.

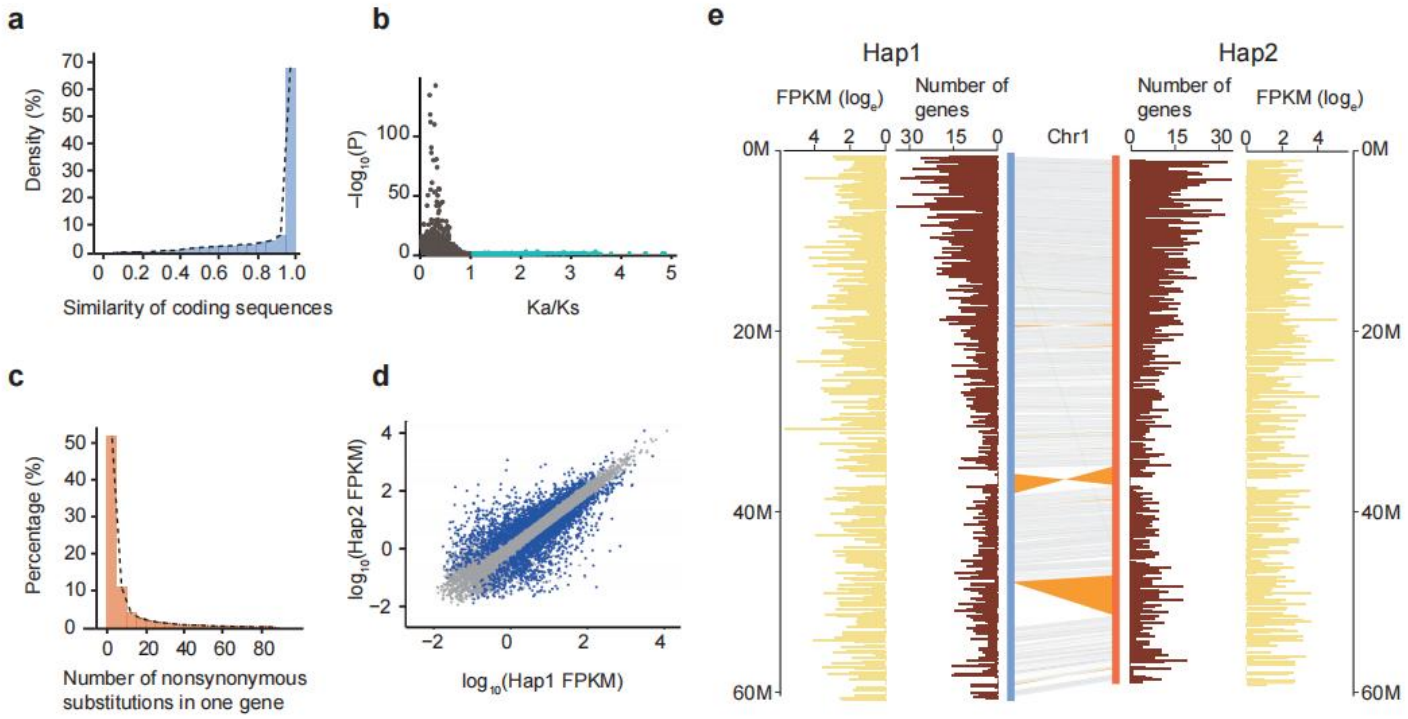

Figure 2

Genetic variations between haplotypes and allelic imbalance in *R. rugosa*. a, Similarities among allele coding sequences. b, Distribution of the Ka/Ks ratio among allelic genes. Blue dots indicate genes with Ka/Ks >1, whereas black dots indicate genes with Ka/Ks <1. *P* values were calculated using a two-sided Fisher's exact test. c, Distribution of the number of non-synonymous substitutions between alleles. d, Identification of ASE genes in leaves. Coordinates are logarithmically scaled (log_{10}). Blue dots represent ASE genes, whereas gray dots represent non-ASE genes. FPKM, fragments per kilobase exon per million fragments mapped; Hap1, haplotype 1; Hap2, haplotype 2. e, Haplotype divergence in a diploid rose genome. The central blue and orange bars represent the two haplotypes of chromosome 1. The gray lines indicate syntenic regions, whereas the orange lines indicate regions with inversions or translocations. Gene numbers and mean FPKM values were determined in 200-kb windows.

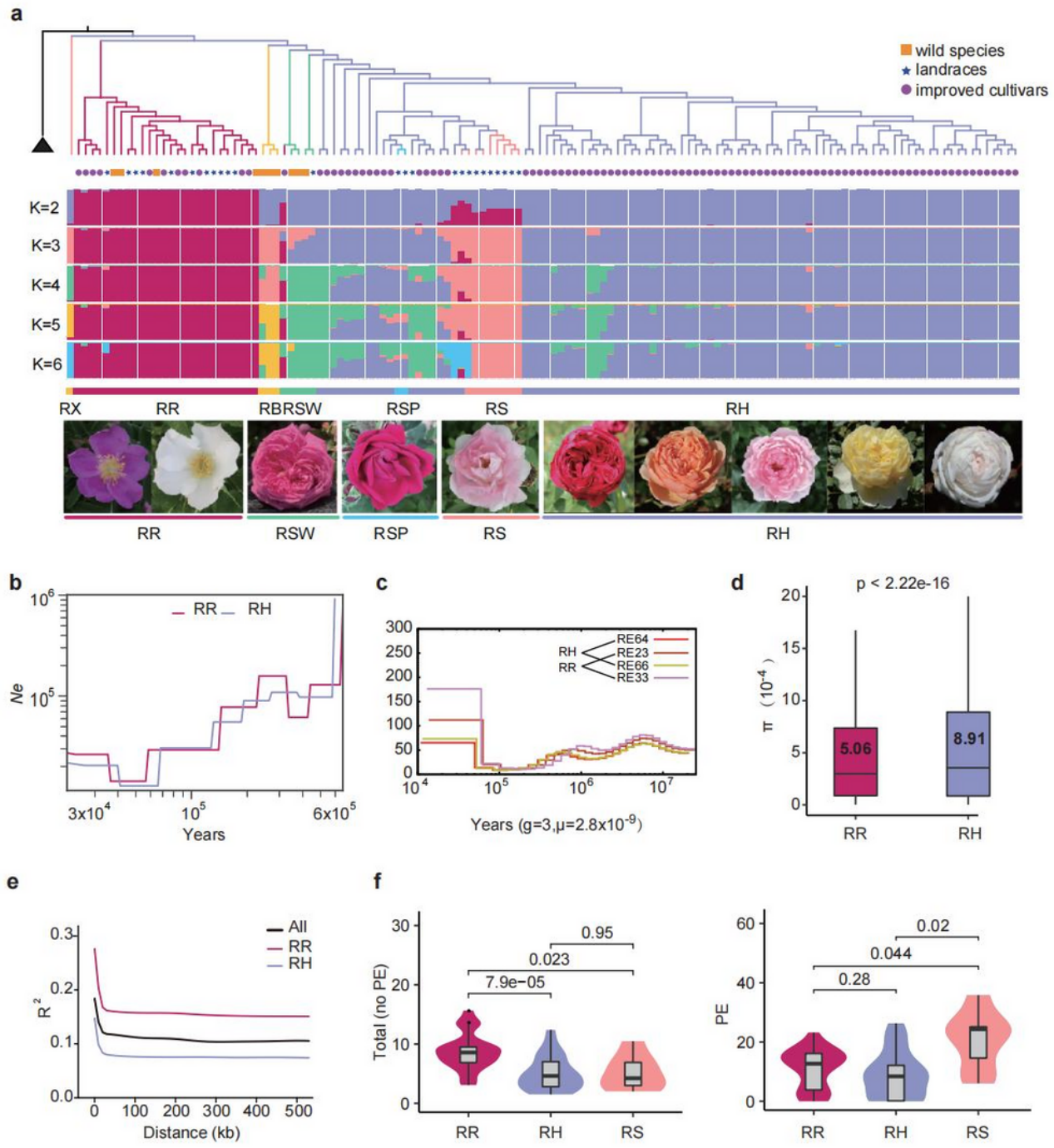

Figure 3

Population structure and genetic divergence of *R. rugosa* (RR), scented *R. hybrida* (RH), and fragrant *Rosa* species (RB, RSW, RSP, RS, and RX). a, Group structure of 133 rose accessions, including 28 *R. rugosa* cultivars, 86 scented *R. hybrida* accessions, and 19 closely related fragrant *Rosa* species. The neighbor-joining tree was constructed and the model-based clustering analysis (K = 2, 3, 4, 5, and 6) was performed using 1,195,460 SNPs. b, c, Demographic analysis of RR and RH. The SMC++ (b) and PSMC (c) models

were used to infer the effective population fluctuations under a mutation rate (μ) of 2.8×10^{-9} per site per generation and a generation time of 3 years. d, Genetic diversity (θπ) for RR and RH. e, Linkage disequilibrium (LD) decay (r^2) in the two groups (RR and RH). f. Aroma substance contents (not including PE) and PE contents in RR, RH, and RS.

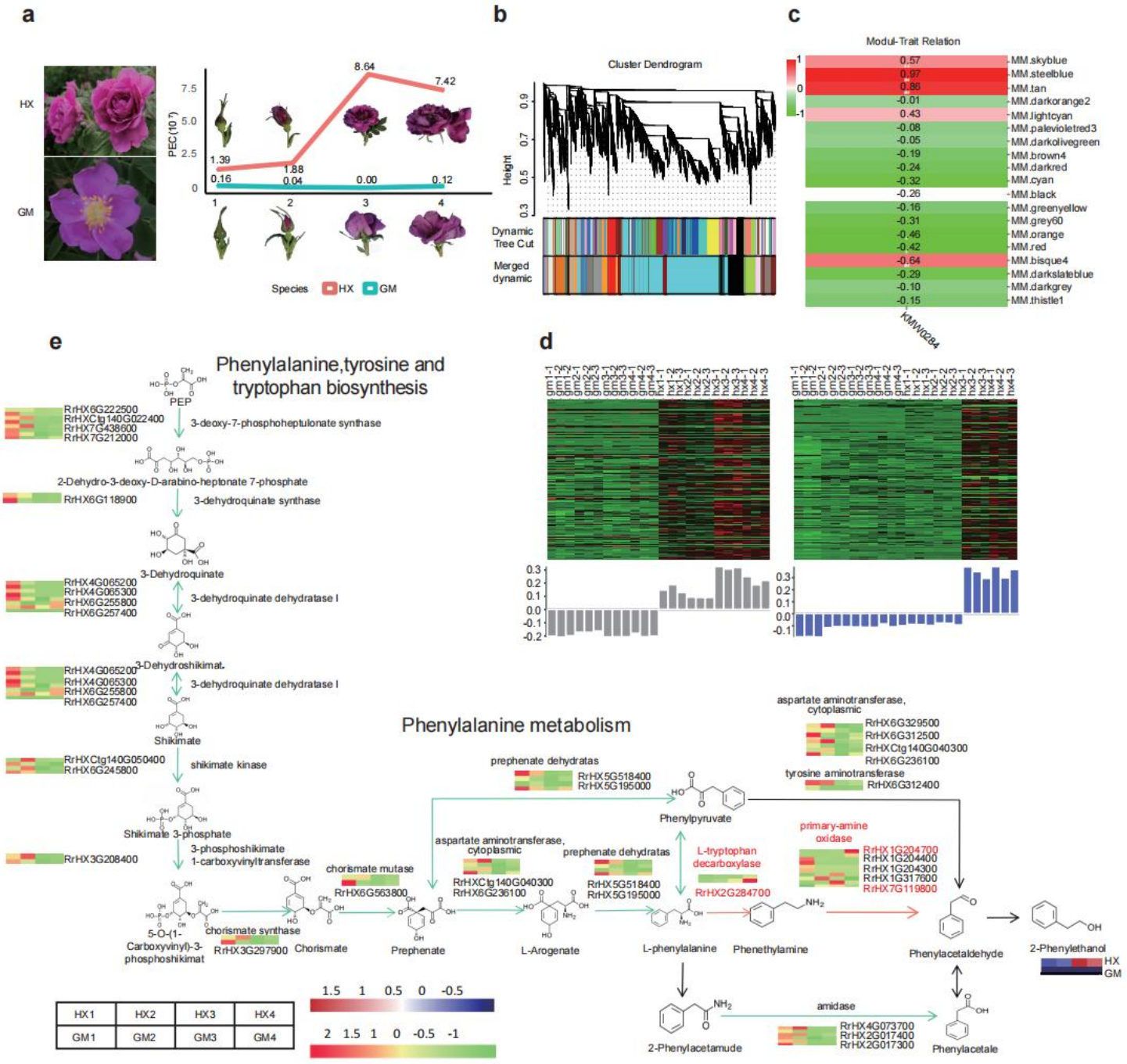

Figure 4

Correlation between the transcriptomes, expression profiles, and transcriptional regulatory networks associated with the modules of two *R. rugosa* varieties (HX and GM) differing regarding fragrance type at four flower development stages. a, Two *R. rugosa* varieties (HX and GM) with significantly different PE contents (PECs) at four flower development stages. b, Module hierarchical clustering tree. Dynamic Tree Cut presents the module division based on the clustering results, whereas Merged dynamic presents the module division. c, Module–PEC relationship diagram. The number in each grid represents the correlation between the module and the PEC, with the *P*-value provided in parentheses. d, Tan and steelblue modules, which comprised genes that were more highly expressed in the HX petals than in the GM petals. Heat maps present the expression profiles of the co-expressed genes (sample names on top) in the modules (labeled on top). Red and green represent up-regulated and down-regulated expression levels, respectively. Bar graphs (below the heat maps) present the consensus expression patterns of the co-expressed genes in each module. e, Expression profile and transcriptional regulatory network associated with the modules with the opposite expression patterns in HX and GM at each flower development stage.

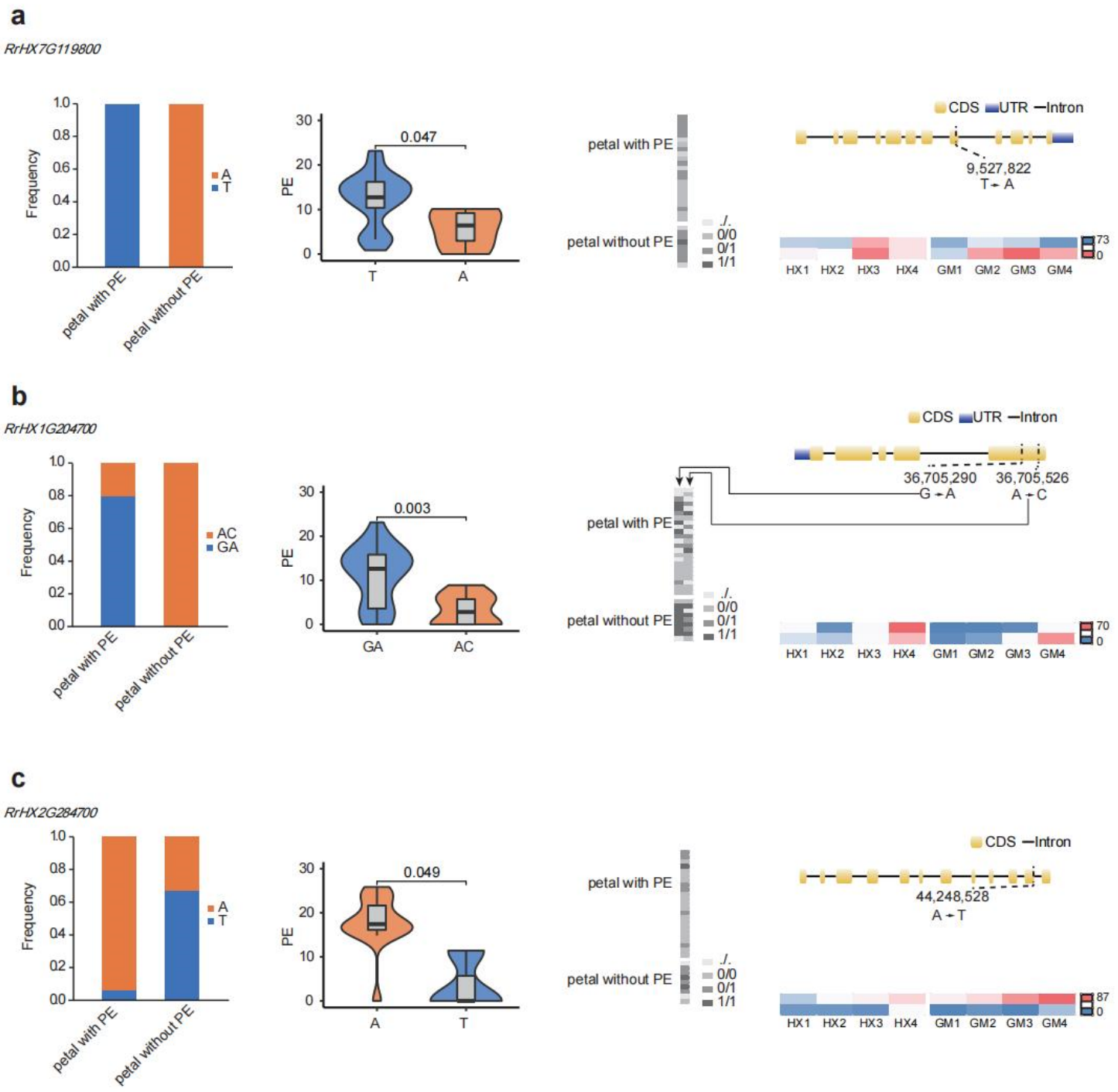

Figure 5

Variations in two fragrance-related genes in the population and in other genomes. a, *RrHX7G119800* gene structure and non-synonymous mutation sites. From left to right: haplotype frequency changes for *RrHX7G119800* in different groups, PE content for haplotypes T and A, genotypic distribution due to one non-synonymous SNP variation ('/' represents deletion, 0/0 represents genotype TT, 0/1 represents genotype TA, and 1/1 represents genotype AA), and *RrHX7G119800* with different allelic expression patterns in four flowering periods, indicative of the differences in the allele functions during different developmental stages. b, *RrHX1G204700* gene structure and non-synonymous mutation sites. From left

to right: haplotype frequency changes for *RrHX1G204700* in different groups, PE content for haplotypes GA and AC, genotypic distribution due to two non-synonymous SNP variations, and *RrHX1G204700* with different allelic expression patterns during four flowering periods. c, *RrHX2G284700* gene structure and non-synonymous mutation sites. From left to right: haplotype frequency changes for *RrHX2G284700* in different groups, PE content for haplotypes A and T, genotypic distribution due to one non-synonymous SNP variation, and *RrHX2G284700* with different allelic expression patterns during four flowering periods.

Supplementary Files

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

- Supplementfigures.docx
- SupplementaryTables.xlsx