DNA Research, 2021, 28(5), 1–9
doi: 10.1093/dnares/dsab017
Advance Access Publication Date: 9 September 2021
Resource Article: Genomes Explored

Resource Article: Genomes Explored

A high-quality chromosome-level genome of wild *Rosa rugosa*

Fengqi Zang[1†], Yan Ma[2†], Xiaolong Tu[3], Ping Huang[1], Qichao Wu[2], Zhimin Li[3], Tao Liu[3], Furong Lin[1], Surui Pei[3], Dekui Zang[2], Xuemei Zhang[3], Yongqi Zheng[1]*, and Yunyan Yu[2]*

[1]State Key Laboratory of Tree Genetics and Breeding; Key Laboratory of Forest Silviculture and Tree Cultivation, National Forestry and Grassland Administration, Research Institute of Forestry, Chinese Academy of Forestry, Beijing 100091, P.R. China, [2]Key Laboratory of State Forestry Administration for Silviculture of the Lower Yellow River, College of Forestry, Shandong Agricultural University, Tai'an 271018, P. R. China, and [3]Annoroad Gene Technology (Beijing) Co., Ltd, Beijing 100176, P. R. China

*To whom correspondence should be addressed. Email: zyq8565@126.com (Y. Z.); yxyxst20040214@163.com (Y. Y.). Tel: +86-10-62888565 (Y. Z.); +86-538-8242602 (Y. Y.). Fax: +86-10-62888565 (Y. Z.); +86-538-8249164(Y. Y.).

†These authors contributed equally to this work and should be considered co-first authors.

Received 8 April 2021; Editorial decision 20 August 2021

Abstract

Rosa rugosa is an important shrub with economic, ecological, and pharmaceutical value. A high-quality chromosome-scale genome for *R. rugosa* sequences was assembled using PacBio and Hi-C technologies. The final assembly genome sequences size was about 407.1 Mb, the contig N50 size was 2.85 Mb, and the scaffold N50 size was 56.6 Mb. More than 98% of the assembled genome sequences were anchored to seven pseudochromosomes (402.9 Mb). The genome contained 37,512 protein-coding genes, with 37,016 genes (98.68%) that were functionally annotated, and 206.67 Mb (50.76%) of the assembled sequences are repetitive sequences. Phylogenetic analyses indicated that *R. rugosa* diverged from *Rosa chinensis* ~6.6 million years ago, and no lineage-specific whole-genome duplication event occurred after divergence from *R. chinensis*. Chromosome synteny analysis demonstrated highly conserved synteny between *R. rugosa* and *R. chinensis*, between *R. rugosa* and *Prunus persica* as well. Comparative genome and transcriptome analysis revealed genes related to colour, scent, and environment adaptation. The chromosome-level reference genome provides important genomic resources for molecular-assisted breeding and horticultural comparative genomics research.

Key words: *Rosa rugosa*; genome sequencing; Hi-C assembly; genome annotation; chromosome synteny.

1. Introduction

Rosa rugosa, belonging to *Rosa* section Cinnamomeae in the Rosaceae family, is a deciduous shrub (Fig. 1a). It is an economically, ecologically, and medicinally important plant with features such as strong resistance to drought and barren land, cold endurance, and wide adaptation to harsh natural environments. The essential oil extracted from roses is of significant economic value and is called "liquid gold". Being widely used in the production of high-grade perfume and cosmetics, rose essential oil is an irreplaceable raw material in the global perfume industry.

Rosa rugosa has been planted for ages for its horticultural and medicinal value in China, mainly in cultivation areas including Pingyin County in Shandong Province, Mt. Miaofengshan in Beijing city, and Kushui town in Gansu Province. Thirty-six cultivars are

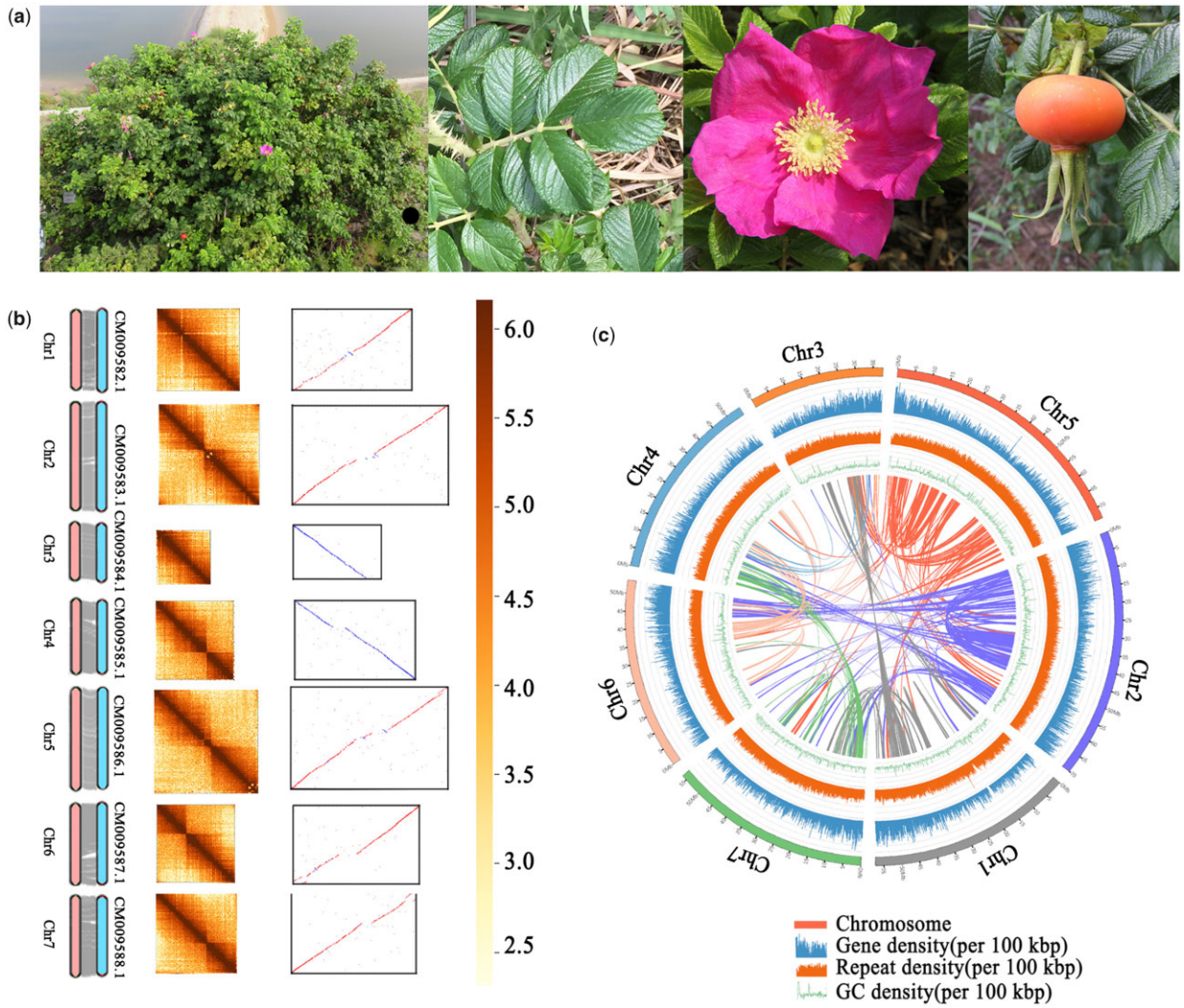

Figure 1. The basic morphology of rugose rose and the assembly and annotation of its genomes. (a) Morphological characteristics of *R. rugosa*. The whole shrub, compound leaf, flowers (3 days post-flowering), and fruits. (b) *Rosa rugosa* genome assembly quality. The images are sized according to relative chromosome size. Chromosome synteny blocks of *R. rugosa and R. chinensis* (left charts). The contact probability on each chromosome is shown on the Hi-C chromosome contact map. The darker the colour, the greater the probability of contact (middle charts). Gene collinearity between the genome of *R. rugosa and R. chinensis* (right charts). (c) Distribution of basic genomic elements of *R. rugosa*.

grown in Pingyin County.[1] Many forms, such as *R. rugosa* f. *rosea*, f. *alba*, f. *plena*, and f. *albo-plena*, have been developed via long-term cultivation.[2] *Rosa rugosa* was introduced to Europe in 1796 and then introduced to the USA in the 19th century. Since then, the cultivation of ornamental varieties has vigorously developed. Many new varieties were developed by crossing *R. rugosa* with *Rosa odorata*, hybrid tea rose, hybrid rose, and other rose species.[3]

Rosa rugosa originated in East Asia and usually grows on coastal hillsides, in sandy soils on seashores, and on offshore islands of E Jilin Province, Liaoning Province and NE Shandong Province in China, where it has become an endangered wild plant because of picking and uprooting.[4–6] However, it has become a naturalized plant in many countries, such as the USA and the Netherlands.[7,8]

Rosa is one of the most important genera in Rosaceae. More than 200 species have been described and are widely distributed from subtropical to cold-temperate regions.[9] Several species and hybrids are cultivated and widely utilized, such as *R. centifolia*, *R. damascena*, *R. gallica*, *R. chinensis*, *R. odorata*, and *R. multiflora*. In recent years, some researchers have sequenced and assembled the genome sequences of *R. chinensis* cultivar Old Blush and *R. multiflora*.[10,11] In order to deepen the basic understanding of *R. rugosa* and its relatives, it is necessary to analyze the whole genome of *R. rugosa*.

According to our research, a *R. rugosa* genome sequence of high quality is reported here, obtained by combining a long-read Pacific Biosciences (PacBio) single-molecule real-time (SMRT) sequencing approach with the Illumina HiSeq X Ten and NovaSeq platforms. This study has important guiding significance for comparative genomics, gene function research and molecular-assisted breeding of Rosaceae in the future.

2. Materials and methods

2.1. Plant materials

The wild *R. rugosa* material studied here was obtained from Yantai, Shandong (37.465° N, 121.695° E). We only used young leaves for

genome library preparation. Besides, we collected samples from five different tissues of healthy individuals for transcriptome sequencing: young leaves, petals, young fruits, mature fruits and receptacles in April, June, and September 2019. All of these tissues were rapidly frozen in liquid nitrogen, later maintained at −80 °C until RNA and DNA were extracted.

2.2. Genome sequencing

We adopted the delicate leaves from the same individual to make DNA extraction by the DNeasy Plant Mini Kit (QIAGEN, CA, USA). To ascertain the purity of the genomic DNA and the concentration of the DNA, we observed genomic DNA in 1% agarose gel and employed a NanoPhotometer instrument (Implen, CA, USA) and a Qubit2.0 Fluorometer (Life Technologies, CA, USA). After obtaining high-quality purified genomic DNA samples, 350 bp short insert libraries were provided from 5 μg whole-genome DNA to do Illumina sequencing. Every library was sequenced adopting PE150 pairing on Illumina HiSeq X Ten platform, and paired-end reads were obtained.

A SMRTbell library was constructed for PacBio sequencing via the PacBio Sequel II sequencing platform following standard PacBio protocols at Annoroad Gene Technology Co., Ltd, (Beijing, China). In brief, 8 μg of high-quality genomic DNA was sheared to fragment and then subjected to damage repair, end repair, adapter ligation, and size selection. Eventually, the DNA was loaded onto the PacBio Sequel II system to read the sequences of the templates (https://www.pacb.com/blog/award-winning-sequel-ii-system/).

2.3. Transcriptome sequencing

Total RNA from five tissues was isolated using the phenol/chloroform method, and purity was detected by a NanoDrop 2000 microspectrometer (Implen, CA, USA).[12]

RNA integrity was detected adopting Agilent Bioanalyzer 2100 RNA Nano 6000 Assay Kit (Agilent Technologies, Santa Clara, CA, USA) (Supplementary Table S1). The RNA Integrity Number (RIN) of the samples was 7.8–10, satisfied for RNA-Seq library construction. Five sequencing libraries from five tissues were constructed by Illumina standard mRNA-seq prep kit and sequenced on an Illumina NovaSeq 6000 with the PE150 mode.

2.4. Estimation of genome size, heterozygosity, and repeat content

The genome size, heterozygosity, and repeat content of *R. rugosa* were estimated by k-mer frequency.[13] A 17-mer frequency was generated from Illumina clean paired-end reads by JELLYFISH (v2.2.0) (http://www.cbcb.umd.edu/software/jellyfish/) with default parameters. The distribution of 17-mers following a Poisson's distribution can reflect the characteristics of the *R. rugosa* genome.

2.5. Genome sequences assembly and quality evaluation

Subreads exported from Sequel II are evaluated quality using the built-in high-quality region finder (HQRF), which recognized the longest high-quality region of each read sequence formed by a single DNA polymerase based on its signal-to-noise ratio.

The *R. rugosa* genome sequences was de novo assembled based on PacBio long reads using Canu v1.8[14] with default parameters. The Redundans pipeline was used to detect and selectively remove redundant contigs and generate a non-redundant draft genome.[15] The consensus genome was exposed to a final round of base-error correction (polishing) using the Illumina reads with BWA (v0.7.9a) and Pilon (v1.22).[16] The completeness of assembled *R. rugosa* genome sequences was assessed using BUSCO (v3.0.1) based on embryophyta_odb10 (issued 2020-08-05, including 1,614 protein).[17] Illumina short reads were mapped against the assembled genome sequences to evaluate the genome coverage based on reads mapping rates. GC content distribution helped to check sample contamination.

2.6. Pseudochromosome construction using Hi-C technology

The Hi-C library was constructed according to the standard procedure.[18] Nuclear DNA of *Rosa* was crosslinked in situ with formaldehyde, extracted, and digested with restriction enzyme *MboI*. The sticky ends of the digested fragments were biotinylated, diluted and ligated randomly. The samples were assessed for quality and biotinylated DNA fragments were enriched and sheared to a fragment size of 100–500 bp by sonication to construct a sequencing library. After A-tailing, pulldown, and adapter ligation, the DNA library was sequenced on an Illumina NovaSeq 6000 in PE150 mode.

The clean Hi-C reads were primarily mapped to the genome employing Bowtie 2 (v2.2.3) by setting the parameters to "–very-sensitive -L 30 –score-min L,-0.6,-0.2 –end-to-end –reorder –rg-id BMG—phred33-quals -p 5".[19] Following the principle of the Hi-C approach, Hi-Pro (v2.7.8) was used to process the mapped Hi-C reads to obtain valid Reads pairs and generated normalized contact maps.[20]

Efficient interaction pairs were used to construct interaction matrices and extend proportionally the initial genome sequences assembled contigs to chromosome-scale scaffolds (hereinafter referred to as pseudochromosome) by LACHESIS.[21] First, the agglomerative hierarchical clustering algorithm was used to cluster the chromosomal groups, and then the contigs of each chromosomal group are ordered and oriented to become pseudochromosomes. The contigs were ordered by LACHESIS through framing a graph and picking up the longest path as the trunk (the highest-confidence order of the contigs in a chromosomal group). Contigs excluded from the trunk are reinserted based on the link information between neighboring contigs. A weighted, directed and acyclic graph (WDAG) was built to orient the ordered contigs.[22] Here, we set CLUSTER_N = 7 for LACHESIS.

Heatmap was constructed to test the accuracy based on the interaction signals.

2.7. Genome annotation

Genome annotation is the basis of current functional genomics research. For *R. rugosa* repeat sequence prediction, two strategies were used. Homology-based prediction was performed using RepeatMasker[23] and RepeatProteinMask software,[24] according to search against Repbase database.[23] We used RepeatModeler (http://www.repeatmasker.org/RepeatModeler/) for *ab initio* prediction. We established a *de novo* repeat sequence library first based on a self-alignment. And then repeat sequences were predicted by RepeatMasker software with default parameters. We adopted Tandem Repeats Finder (TRF) to explore tandem repeats. Two strategies were employed for non-coding RNA prediction, sequence homology search and structure prediction. For rRNAs, snRNAs, miRNAs, the sequences were homology searched against the known non-coding RNA libraries in Rfam (http://rfam.xfam.org/).[25] tRNA

prediction was performed using tRNAscan-SE software[26] based on their structures.

Gene structure prediction was accomplished using the transcriptional data of *R. rugosa*, homolog-based prediction and *de novo* prediction. Trinity was used to assemble RNA-seq, the assembled transcripts were mapped to the genome using GMAP (http://research-pub.gene.com/gmap) and PASA (http://pasa.sourceforge.net/) was employed to predict gene models. Protein-coding sequences of known homologs in five species, *Fragaria vesca* (GenBank assembly accession: GCA_000184155.1), *Malus domestica* (GenBank assembly accession: GCA_004115385.1), *Prunus mume* (GenBank assembly accession: GCA_000346735), *Prunus persica* (GenBank assembly accession: GCA_000346465.2) and *R. chinensis* (GenBank assembly accession: GCF_002994745.1), were aligned to the genome of *R. rugosa* by tblastn (E-value cutoff: 1e-5), and gene structure was predicted by Genewise v2.2.0 (http://www.ebi.ac.uk/~birney/wise2/). *Ab initio* prediction was performed by Augustus v3.3,[27] GeneMark v4.33,[28] and SNAP (https://github.com/KorfLab/SNAP). EVidenceModeler (EVM) was used to integrate the above forecast results into a non-redundant and high confidence gene set.[29]

Functional annotation of the protein-coding genes was implemented employing BLAST against the following public databases: UniProt (https://www.uniprot.org/), NT (https://www.ncbi.nlm.nih.gov/nucleotide/), NR (ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz), EggNOG,[30] GO,[31] KEGG,[32] and HMMER (v3.1) against Pfam.[33]

2.8. Comparative phylogenomics

The protein sequences from *R. rugosa*, *M. domestica*, *P. persica*, *F. vesca*, *Pyrus bretschneideri*, *R. chinensis*, *P. mume*, *Vitis vinifera*, *Arabidopsis thaliana*, and *Oryza sativa* were used for gene families clustering. Genes encoding less than 50 amino acids were removed, the longest transcript were kept for subsequent analysis. An all-vs-all comparison using BLAST-2.2.26 was performed with the following parameters: "p blastp -m 8 -e 1e-5 -a 10 -F F". OrthoMCL [34] was used to classify gene families that were potentially orthologs, inparalogs, or co-orthologs. Muscle (v3.8.31) was used to align the orthologs.[35] Grounded on 882 single-copy genes, a ten-species phylogenetic tree was constructed by using PhyML (v 3.0) with "-d nt -b -4 -m HKY85 -a e -c 4 -t e".[36] MCMCTree of PAML (v4.9)[37] was adopted to estimate divergence times between the sampled species with "clock = correlated rates, model = JC69, burnin = 20,000, nsample = 100,000", the calibration time was obtained from the TimeTree database (http://www.timetree.org/).[38] Expansion and contraction of the orthologous gene families were identified using CAFÉ (v 4.1).[39] To investigate the evolution of *R. rugosa* chromosomes, we looked for protein sequences in *R. rugosa* that adopted blastp (E < 1e-5) to identify collinear blocks. In addition, the protein sequences from *R. rugosa* were searched against those of *P. persica*, *F. vesca*, *R. chinensis*, and *A. thaliana*. The results were analyzed with MCscan (v0.8) [40] with default parameters to identify syntenic blocks. Via anchoring the aligned *R. rugosa* genes to the peach genome and rose genome, gene synteny was obtained. We then calculated the 4DTv (the transversion rate at fourfold degenerate third-codon positions) values for all gene pairs in syntenic blocks and plotted their distribution.

2.9. GO and KEGG enrichment analyses

GO and KEGG enrichment analyses were assigned using Blast2GO v4.1.9 with default parameters and the KEGG database (https://www.genome.jp/kegg/).

3. Results and discussion

3.1. Genome sequencing and assembly

In total, ~417.54 million raw reads were generated, representing ~62.63 Gb (~150× the assembled genome sequences). *K*-mer analysis revealed that the genome size was 376.56 Mb, the heterozygosity was 1%, and the repeat content was 54.1% (Table 1, Supplementary Table S2 and Fig. S1). Heterozygosity of *R. rugosa* was higher than that of some plants of Rosaceae, such as *P. armeniaca* (about 0.09%) [42] and *P. mume* (about 0.03%),[43] but close to *P. bretschneideri* (about 1.02%) [44] and the released *R. rugosa* genome (0.71%).[45] Using the rice cultivar Nipponbare and maize B73 as internal references, flow cytometry analysis showed that the genome size of *R. rugosa* was about 0.5 Gb. Owing to natural autoincompatibility and recent interspecific hybridization, all roses have highly heterozygous genomes that are challenging to assemble.[41]

Amount to 7.9 million PacBio subreads were generated, with sequencing data of ~131.97 Gb. The average read length was 16.7 kb, and the N50 was 24.5 kb. We used Canu to assemble the initial genome of 766 Mb, which was nearly twice the size of the estimated genome. Finally, the genome size of *R. rugosa* was 407 Mb and contig N50 was 2.85 Mb, after removing redundancy sequence, closed to the released *R. rugosa* genome (382.64 Mb) [45] Genome assembly was larger than expected, mainly because of genome high hybridization. Approximately 96.91% of the Illumina PE reads was mapped to the assembled genome sequences, covering 91.6% of assembly sequences (Supplementary Table S3). The genome assembly completeness was assessed using BUSCO. In total, we identified 1,503 (93.2%) complete BUSCOs and 10 (0.6%) fragmented BUSCOs in *R. rugosa* genome, similar to the released *R. rugosa* genome (93.2%),[45] indicating high genome assembly quality (Supplementary Table S4).

3.2. Pseudochromosome construction

We adopted an Illumina NovaSeq 6000 PE150 platform to sequence a Hi-C DNA library, in total, we get the clean data of 46.92 Gb size (~115× of assembled genome size).

These data were mapped to the assembled contigs, and ~20 million valid paired-end reads were used to build the pseudochromosomes with LACHESIS. Finally, seven pseudochromosomes were assembled, which covered 98.96% (~402.9 Mb) of the genome assembly (~407.1 Mb), keeping the characteristic of 350 contigs and 62 scaffolds (with a contig N50 of 2.85 Mb and a scaffold N50 of 56.6 Mb). The maximum length of the pseudochromosomes was 73.25 Mb, and the minimum length was 37.69 Mb (Table 2, Supplementary Tables S5 and S6, Fig. 1b). The pseudochromosomes were numbered based on the syntenic relationship with *R. chinensis*. Comparison of genomic assemblies of *R. rugosa*, *R. chinensis*, and

Table 1. Statistics of *R. rugosa* genome size, heterozygosity, and repeat ratio

Sample	*R. rugosa*
K-mer	17
K-mer number	55,022,631,402
K-mer depth	143
Genome size (Mbp)	369.56
Revised genome size (Mbp)	376.56
Heterozygous ratio (%)	1
Repeat (%)	54.10

F. vesca indicated that *R. rugosa* genome assembly was high quality, providing an opportunity for comprehensive evaluation of genomic variation of Rosaceae (Supplementary Table S7).

3.3. Genome annotation

The predicted repeat sequences represented 206.67 Mb (50.76%) of the *R. rugosa* genome assembly. We found that the largest part of the repeat sequences is the retrotransposon (class I elements), Similar to the pattern in *R. chinensis* (HapOB) and *F. vesca* (Table 3), (class I elements) LTR retrotransposons were the most abundant repeat elements in *R. rugosa*, represented 78.72 Mb (19.34%) of the total repeats. DNA transposons accounted for only 13 Mb (3.24%) of all the combined repeats (Table 3).

In the preliminary comparison, more repeat contents were found in *R. rugosa* (49.84%) and *R. chinensis* (63.21%) than in *F. vesca* (35.44%). This finding suggests that the differences in genome size among *R. rugosa*, *R. chinensis*, and *F. vesca* are due to the expansion of the transposable element, especially LTR retrotransposons expansion (Supplementary Table S8). To make accurate predictions of the coding gene models, and we developed a large RNA-seq data set from five different tissues of wild rose. Approximately 167,174 transcript sequences were assembled. A comprehensive strategy was implemented to integrate *de novo* predictors, protein homology search, and *de novo* assembly transcripts. Integrated, the above forecast results with EVM, 37,512 predicted genes were finally obtained (Table 4), closed to the reported gene numbers in *R. rugosa* (39,704) and *R. chinensis* (36,377).[45]

Table 2. Number of Contigs anchored employing the Hi-C technology

Pseudochromosome	Number of anchored contigs	Sequence length (bp)
chr1	32	57,358,339
chr2	50	70,076,285
chr3	34	37,694,776
chr4	38	54,226,211
chr5	50	73,248,502
chr6	52	53,730,656
chr7	39	56,572,142
Total (ratio %)	84.28	98.96
Total number of anchored contigs	295	402,906,911

We adopted GO, KEGG, Pfam, NT, EggNOG, UniProt, KO, and NR databases to perform functional annotation for all genes that were predicted. Altogether, 37,016 genes accounting for 98.68% of all gene sets were functionally annotated by the different databases (Supplementary Table S9 and Figs S2–S4). The number of predicted proteins in *R. rugosa* (37,512) was higher than that in *F. vesca* (28,588)[46] and similar to that in other *Rosa* species (39,669).[10]

The predicted non-coding genes included 954 miRNAs, 780 tRNAs, 803 rRNAs, and 241 snRNAs with a total length of 503,477 bp (0.12% of the whole-genome length) (Supplementary Table S10).

3.4. Ortholog groups identification and phylogenetic analysis

To investigate the speciation of *R. rugosa*, the correlations of its protein-coding genes were analyzed with those of *M. domestica*, *P. bretschneideri*, *P. persica*, *P. mume*, *F. vesca*, *O. sativa*, *A. thaliana*, *R. chinensis*, and *V. vinifera*. Among the 37,512 protein-coding genes, 30,002 genes were clustered into 17,418 ortholog groups, with an average of 1.72 genes per ortholog group, of which 1366 were unique ortholog groups. (Supplementary Table S11, Fig. 2b).

A phylogenetic tree was constructed using 882 single-copy ortholog groups. The median divergence time of rugose rose and Chinese rose was ~6.6 million years ago (MYA) (4.1–9.2 MYA), and that of *Rosa and F. vesca* from their most recent common ancestor was 19.0 MYA (10.7–32.2 MYA) (Fig. 2a). According to the 4DTv distribution, there was no recent whole-genome duplication (WGD) event after species differentiation between *R. rugosa and R. chinensis* (Fig. 2c) and only one significant group of blocks suggestive of a WGD event in *R. rugosa*, which was the hexaploidy event common in eudicots. The 4DTv values peaked at 0.56. In Rosaceae, *P. persica* [47] and *P. mume*,[43] belonging to subfamily Prunoideae, underwent only one WGD event, similar to *R. rugosa*, with a peak 4DTv value of 0.56. Two WGD events occurred during the speciation of *P. bretschneideri*, containing a recent event with a 0.08–4DTv-value and an ancient event with a 0.5–4DTv-value. *M. domestica and P. bretschneideri*,[46,48] belonging to Maloideae, shared the most recent WGD event. The number of chromosomes in *P. bretschneideri and M. domestica* was $2n = 2x = 34$, which was almost twice that in *R. rugosa* ($2n = 2x = 14$), *P. persica* ($2n = 2x = 16$), and *P. mume* ($2n = 2x = 16$). Recent WGD events did not occur in *P. persica*, *P. mume*, or *R. rugosa*.

Table 3. Repeat sequence prediction in *R. rugosa*

Type		Number	Length (bp)	Fraction of genome (%)
Class I (Retransposons)	Class I/LTR	116,901	78,724,804	19.34
	Class I/LINE	23,368	12,901,450	3.17
	Class I/SINE	58	4,102	0.001
Class II (DNA transposons)	DNA	48,398	13,203,968	3.24
	Class II/Crypton	1,438	96,994	0.02
	Class II/Maverick	548	43,313	0.01
	Other	46,412	13,063,661	3.21
Unknown		244,054	106,441,527	26.14
Other		215,439	11,045,419	2.71
Total with overlap		426,429	202,907,373	49.84

Table 4. Prediction of protein-coding genes

Method	Software	Species	Gene number
Ab initio	GeneMark	—	40,311
	Augustus	—	38,224
	SNAP	—	30,761
Homology-based	Genewise	*Fragaria vesca*	37,692
		Malus domestica	31,178
		Prunus mume	29,098
		Prunus persica	30,951
		Rosa chinensis	53,271
RNA-seq	PASA	—	32,402
Integration	EVM	—	37,512

3.5. Synteny of *R. rugosa* with *R. chinensis and P. persica*

We compared the *R. rugosa* genome with the *R. chinensis and P. persica* genomes to analyze the synteny in detail (Fig. 2d; Supplementary Fig. S5). *Rosa rugosa and R. chinensis* show highly conserved synteny. In Rosaceae, synteny between *P. persica* and *R. rugosa* was also well conserved. According to the orthologous gene orders, a total of 604 and 643 gene blocks were identified, there were 19,068 and 22,924 syntenic gene pairs in *P. persica and R. chinensis* genomes, respectively. The mean numbers of gene pairs per block were 31.57 and 35.65 for the *P. persica and R. chinensis and R. rugosa* genomes, respectively. Large macrosyntenic blocks were conserved between these two species. *Rosa rugosa* chromosomes (Chrs) 1, 2, 3, 4, 5, 6 and 7 displayed strong synteny with *R. chinensis* Chrs 1, 2, 3, 4, 5, 6 and 7, respectively.

Identifying good syntenic chromosome pairs in *R. rugosa* is challenging (Fig. 1c) because of its self-collinearity, in contrast to the rearrangement of chromosomes identified in the genomes of *P. bretschneideri* and *M. domestica*.[45,47] This discrepancy may be due to the recent WGD event did not affect *R. rugosa*. The evolution of the nine ancestral chromosomes of Rosaceae was studied previously. The tripling of the seven chromosomes of Eudicots ancestor may have an additional rearrangement, resulting in the nine ancestral chromosomes of Rosaceae. After the fragmentation and recombination of chromosomes, the ancestors of Rosoideae were differentiated first, and the number of chromosomes was 7. Next, the ancestor of Prunoideae diverged from that of Maloideae. The chromosome number of Prunoideae's ancestor was 8. The chromosomes of Maloideae's ancestors doubled in size once, resulting in 18 chromosomes.

3.6. Unique and expanded ortholog groups

Phylogenetic analyses revealed lineage-specific orthogroups or expanded orthogroups in one genome, generally called unique or expanded ortholog groups. Unique ortholog groups mean lineage-specific orthogroups in one genome, suggested their special functions or pathway in this species. Expanded ortholog groups mean these orthogroups have significantly more gene copies than other species, these genes maybe contributed to specific biological phenotype and important trait in this genome. The unique ortholog groups (3,166) and expanded ortholog groups (1,100) of *R. rugosa* were obtained. To understanding the biological function of these unique and expanded orthogroups in *R. rugosa* genome, GO and KEGG enrichment of these ortholog groups was carried out.

Among the unique ortholog groups, we detected carboxylesterase 18 (CXE18) and dihydrolipoyl dehydrogenase (DLD), two genes that related to pathogen resistance in animals; accelerated cell death 11 (ACD11), which is related to regulating the reactive oxygen species (ROS)-mediated defense response in plants;[49] and protein phosphatase 2C (PP2C), which has been given the notice to participate in responses to abiotic stress (drought, salt tolerance) and seed germination (Supplementary Table S12).[50] These genes are well known for conferring resistance to environmental stresses in plants and are important in defense responses in *R. rugosa*.

We performed GO and KEGG enrichment analyses (Supplementary Figs S6 and S7). The GO enrichment analysis of functional genes showed that "immune response" and "response to nickel cation" were enriched functions for many genes (Supplementary Fig. S6 and Table S13). Thus, these genes may be important for adaptation to adverse environments. The KEGG enrichment analysis showed that the pathway "terpenoid backbone biosynthesis" was enriched, suggesting that these kinds of genes may work in the formation of fragrance (Supplementary Fig. S7 and Table S14).

About 1,110 expanded ortholog groups and 1,361 contracted in *R. rugosa* were identified. In the GO enrichment analysis, some expanded ortholog groups were related to aspects of floral development, such as "auxin transport" and "hormone transport" (Supplementary Table S15 and Fig. S8). We also found some pathways of expanded ortholog groups associated with the regulation of flowering and resistance, suggesting that these ortholog groups may influence flowering time and adaptability to environmental factors, such as cold, bacteria, and heavy metal ions. The KEGG enrichment results revealed genes related to some stress-resistance traits, as evidenced by enriched pathways such as "Cutin, suberin and wax biosynthesis" (Supplementary Table S16 and Fig. S9).

3.7. Flavonoid genes of *R. rugosa*

The colours of *R. rugosa* flowers are usually pink to red. This colouration is attributed to cyanidin or pelargonidin glucosides, which are types of anthocyanins, a class of coloured flavonoids (Supplementary Fig. S10 and Table S17). In higher plants, anthocyanins include a stable form: anthocyanidin 3-glucoside.[51]The *R. rugosa* transcriptome contains genes exhibiting high identity to reported flavonoid biosynthetic enzymes containing chalcone isomerase (CHI), chalcone synthase (CHS), dihydroflavonol 4-reductase (DFR), flavanone 3-hydroxylase (F3H) and flavonoid 3′-hydroxylase (F3′H) (Supplementary Fig. S10). The biological processes of the colours of *R. rugosa* were studied, and it was found that phenylalanine ammonia lyase (PAL), 4-coumarate CoA ligase (4CL), CHI, flavonol synthase (FLS), F3H and F3′H genes were significantly high expressed in leaf (Supplementary Table S18 and Fig. S10). Importantly, F3H and uridine diphosphate glucose (UDP-glucose): anthocyanidin 5, 3-O-glucosyltransferase (A53GT) were also enriched in petals . This is the key enzyme gene for the synthesis of pelargonidin 3, 5-diglucoside and the main cause of the red colour of *R. rugosa* flowers. Moreover, there was no expression detected of flavonoid 3′, 5′-hydroxylase (F3′5′H) and flavone synthase (FNS) in petals, thus, flowers of *R. rugosa* do not contain the flavone or delphinin that are most commonly found in blue or purple flowers. Besides, genes corresponding to FLS and F3′H were also ascertained in rugose rose genome (Supplementary Table S17), while we did not sought out those consistent with F3′5′H and FNS. These results suggest that the lack of blue colour in *R. rugosa* may be as a result of a lack of F3′5′H in

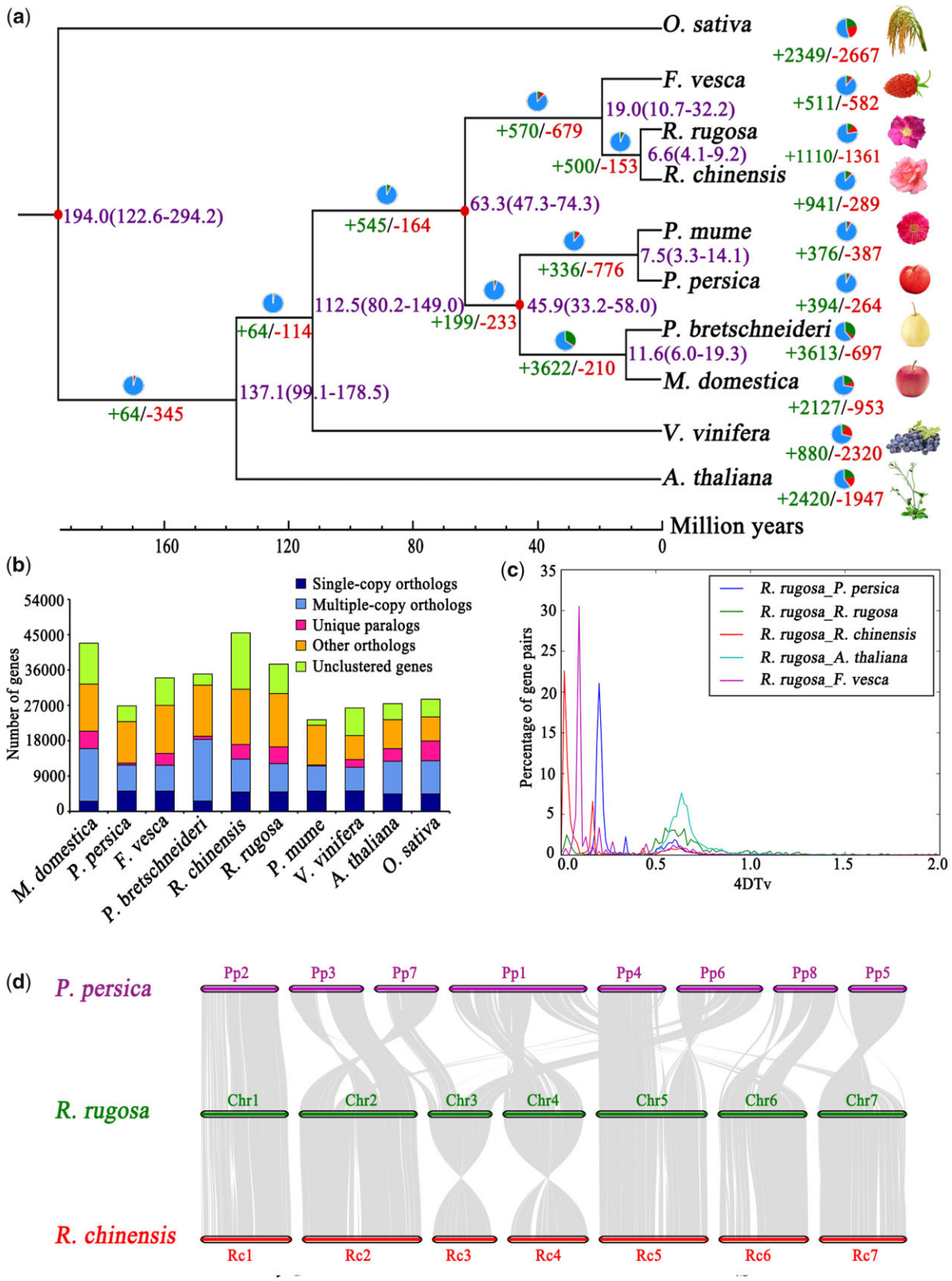

Figure 2. *Rosa rugosa* genome evolution. (a) *Rosa rugosa* diverged from other species and their phylogeny. The purple numbers are the divergence time of the prediction. The numbers below the branches are the numbers of expanded and contracted ortholog groups. The scale at the bottom represents divergence time, and the one-time unit represents 100 million years ago. (b). OrthoMCL clusters of *R. rugosa* and nine other species. (c) Distribution of fourfold degenerate site (4DTv) duplicate gene pair distance in *R. rugosa*, *P. persica*, *R. chinensis*, *A. thaliana*, and *F. vesca*. (d) Chromosome synteny of *R. rugosa and R. chinensis*. Chromosome synteny of *R. rugosa and P. persica*.

the genome; similar results have been found in rose [41] and *R. multiflora*.[11]

3.8. Terpenes and benzenoid biosynthetic genes of *R. rugosa*

The fragrance of roses is one of the most economically valuable characteristics of *R. rugosa*. The main sources of floral fragrances are terpenes, benzene compounds, and fatty acid derivatives. The synthesis of terpenes is divided into the mevalonate pathway (MVA pathway) and methylerythritol phosphate pathway (MEP pathway).[52,53] Citronellol, geraniol, nerol and their acetates, and linalool are the main components of the characteristic fragrance of *R. rugosa*.[54] Sylvie Baudino *et al.* found that the Nudix hydrolase1 (RhNUDX1) is a cytosolic component of a terpene synthase independent pathway for monoterpene biosynthesis that leads to scent production in roses *(Rosa × hybrida)*, which is a special pathway for geraniol synthesis.[55] In addition to terpenes, benzenoid also affects the scent of *R. rugosa*.[56]

The key enzyme genes that participate fully in the terpenes and benzenoid pathways of *R. rugosa* were analyzed. Isopentenyl diphosphate isomerase, 1-deoxy-D-xylulose-5-phosphate synthase, 1-deoxy-D-lxylulose-5-phosphate reductoisomerase, phenylacetaldehyde synthase (PAAS), benzoyl-CoA: benzyl alcohol benzoyl transferase and carotenoid cleavage dioxygenase genes were enriched in the leaf; acetyl-CoA acetyltransferase, geranyl diphosphate synthase, geranylgeranyl diphosphate synthase, PAAS, S-adenosyl-L-methionine, salicylic acid carboxyl methyltransferase, phloroglucinol O-methyltransferase, and S-adenosyl-L-methionine: (iso)eugenol O-methyltransferase genes were enriched in the mature fruit (Supplementary Table S19 and Fig. S11). Additionally, the alcohol acyl transferase (AAT) gene is highly expressed in petals, and as a key enzyme, AAT plays an important role in the last step of catalyzing ester synthesis. Volatile acetate compounds determine some specific fragrances in floral fragrances.[57,58] We also found that RhNUDX1 is highly expressed in petals, maybe there is also a special pathway in *R. rugosa* that regulates geraniol synthesis through RhNUDX1.

4. Conclusion

In this study, a high-quality, chromosomal-scale genome sequence of *R. rugosa* was obtained, with a contig N50 of 2.58 Mb and a scaffold N50 of 56.6 Mb. The assembled genome included 37,512 protein-coding genes, 1,366 unique gene families, 1,100 expanded gene families, and 1,361 contracted gene families. *Rosa rugosa* diverged from its common ancestor with *F. vesca* about 4.1–9.2 MYA. Our results lay the foundation for exploring the special biological features of *R. rugosa* and provide a useful data source for comparative genomics and phylogenomics research among Rosaceae taxa. This genomic information can help to identify important genes related to underlying horticultural traits and accelerate genetic studies and breeding programs for Rosaceae plants.

Supplementary data

Supplementary data are available at DNARES online.

Authors' contributions

Y.Q.Z., F.Q.Z., Y.M. designed the experiment; F.Q.Z., Y.M., P.H., Q.C.W., F.R.L. collected samples and extracted the genomic DNA from samples; F.Q.Z., Y.M., X.L.T., S.R.P., X.M.Z., Z.M.L., D.K.Z worked on the sequencing and data analysis; F.Q.Z., Y.M. wrote the manuscript; Y.Q.Z., Y.Y.Y revised the manuscript.

Acknowledgements

This study was invested by the Shandong Agricultural Seeds Engineering Project (2019LZGC01802), National Natural Science Foundation of China (31870688), and the National Center for Forestry and Grassland Genetic Resources (NCFGGR-2020). There was no additional external funding received for this study.

Accession numbers

The whole-genome sequence data reported in this paper have been deposited in National Genomics Data Center, Beijing Institute of Genomics (BIG), Chinensis Academy of Sciences under Bioproject accession number PRJCA002607 and Biosample accession number PRJCA002607. The genome assembly and gene annotation have been stored in the Genome Warehouse at BIG, under accession number GWHALOL00000000.

Conflict of interest

The authors declare no conflicts of interest.

References

1. Li, Y.S. 2006, *Studies on Germplasm Resource and Cultivars Classification of Rosa rugosa in China*. Beijing Forestry University Press: Beijing.
2. Ku, T.C. 1985, *Rosa rugosa*. In: Yu, T.T.(editor), *Flora Reipublicae Popularis Sinica* . Science Press: Beijing, Vol. 37, pp. 401.
3. Liu, Q.H. 1998, *Studies on Systematic Classification of Rosa rugosa Cultivars in China*. Beijing Forestry University Press: Beijing.
4. Fu, L.G. and Jin, J.M. 1991, *China Plant Red Data Book - Rare and Endangered Plants*. Science Press: Beijing.
5. Zang, D.K. 2016, *Rare and Endangered Plants in Shandong*. China Forestry Publishing House: Beijing.
6. Jiang, L.Y. and Zang, D.K. 2017, Analysis of genetic relationships in *Rosa rugosa* using conserved DNA-derived polymorphism markers, *Biotechnol. Biotechnol. Equip*., **32**, 188–94.
7. Andersen, U.V. 1995, Invasive aliens: a threat to the Danish coastal vegetation. In: Healy, M.G. and Doody, J.P. (eds). *Directions in Europe: A Coastal Management*. Samara Press: Cardigan, pp. 335–344.
8. Hans, H.B. 2005, *Rosa rugosa* Thunb. ex Murray, *J. Ecol*., **93**, 441–70.
9. Ku, T.C. and Kenneth, R.R. 2010, *Rosa rugosa*. In: Wu, Z.Y., Raven, P.H. and Hong, D.Y. (eds). *Flora of China* . Science Press: Beijing; Missouri Botanical Garden Press: St. Louis, MO, vol. 9, pp. 339..
10. Hibrand, S.O., L., Ruttink, T., Hamama, L., et al. 2018, A high-quality genome sequence of *Rosa chinensis* to elucidate ornamental traits, *Nat. Plants*, **4**, 473–84.
11. Nakamura, N., Hirakawa, H., Sato, S., Otagaki, S., Matsumoto, S., Tabata, S. and Tanaka, Y. 2018, Genome structure of *Rosa multiflora*, a wild ancestor of cultivated roses, *DNA Res*., **25**, 113–21.
12. Chomczynski, P. and Sacchi, N. 2006, The single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction: twenty-something years on, *Nat. Protoc*., **1**, 581–5.
13. Liu, B.H., Shi, Y.J. and Yuan, J.Y. 2013, Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects, *Quant. Biol*., **35**, 62–7.

14. Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. and Phillippy, A.M. 2017, Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation, *Genome Res*., **27**, 722–36.
15. Pryszcz, L.P. and Gabaldón, T. 2016, Redundans: an assembly pipeline for highly heterozygous genomes, *Nucleic Acids Res*., **44**, e113.
16. Walker, B.J., Abeel, T., Shea, T., et al. 2014, Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement, *PLoS ONE*, **9**, e112963.
17. Simão, F.A., Waterhouse, R.M., Panagiotis, I., Kriventseva, E.V. and Zdobnov, E.M. 2015, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, *Bioinformatics*, **31**, 3210–2.
18. Lajoie, B.R., Dekker, J. and Kaplan, N. 2015, The Hitchhiker's guide to Hi-C analysis: practical guidelines, *Methods*, **72**, 65–75.
19. Langmead, B. and Salzberg, S.L. 2012, Fast gapped-read alignment with Bowtie 2, *Nat. Methods*., **9**, 357–9.
20. Nicolas, S., Lajoie, B.R., Nora, E.P., et al. 2012, HiTC: exploration of high-throughput 'C' experiments, *Bioinformatics*, **21**, 2843–4.
21. Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O. and Shendure, J. 2013, Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions, *Nat. Biotechnol*., **31**, 1119–25.
22. Nishikawa, K. and Tominaga, N. 2001, Isolation, growth, ultrastructure, and metal tolerance of the green alga, *Chlamydomonas acidophila* (Chlorophyta), *Biosci. Biotechnol. Biochem*., **65**, 2650–6.
23. Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. 2005, Repbase Update, a database of eukaryotic repetitive elements, *Cytogenet. Genome Res*., **110**, 462–7.
24. Tarailo-Graovac, M. and Chen, N. 2009, Using RepeatMasker to identify repetitive elements in genomic sequences, *Curr Protoc. Bioinform*., **25**, Chapter Unit 4 10.
25. Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R. and Bateman, A. 2005, Rfam: annotating non-coding RNAs in complete genomes, *Nucleic Acids Res*., **33**, D121–D124.
26. Lowe, T.M. and Eddy, S.R. 1997, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, *Nucleic Acids Res*., **25**, 955–964.
27. Mario, S., Rasmus, S., Stephan, W. and Burkhard, M. 2004, AUGUSTUS: a web server for gene finding ineukaryotes, *Nucleic Acids Res*., **32**, 309–12.
28. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y.O. and Borodovsky, M. 2008, Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training, *Genome Res*., **18**, 1979–90.
29. Haas, B.J., Delcher, A.L., Mount, S.M., et al. 2003, Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies, *Nucleic Acids Res*., **31**, 5654–66.
30. Powell, S., Szklarczyk, D., Trachana, K., et al. 2012, eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges, *Nucleic Acids Res*., **40**, D284–D289.
31. Ashburner, M., Ball, C.A., Blake, J.A., et al. 2000, Gene ontology: tool for the unification of biology, *Gene*, **25**, 25–9.
32. Kanehisa, M., Goto, S., Sato, Y., Miho, F. and Mao, T. 2012, KEGG for integration and interpretation of largescale molecular data sets, *Nucleic Acids Res*., **40**, D109–D114.
33. Finn, R.D., Alex, B., Jody, C., et al. 2014, Pfam: the protein families database, *Nucleic Acids Res*., **42**, D222–D230.
34. Li, L., Stoeckert, C.J.S. and Roos, D.S. 2003, OrthoMCL: identification of ortholog groups for eukaryotic genomes, *Genome Res*., **13**, 2178–89.
35. Edgar, R.C. 2004, MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res*., **32**, 1792–7.
36. Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W. and Gascuel, O. 2010, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0, *Syst. Biol*., **59**, 307–21.
37. Yang, Z.H. 2007, PAML 4: phylogenetic analysis by maximum likelihood, *Mol. Biol. Evol*., **24**, 1586–91.
38. Conesa, A. and Gotz, S. 2008, Blast2GO: a comprehensive suite for functional analysis in plant genomics, *Int. J. Plant Genomics*, **4**, 619832.
39. Bie, T.D., Cristianini, N., Demuth, J.P., et al. 2006, CAFE: a computational tool for the study of gene family evolution, *Bioinformatics*, **22**, 1269–71.
40. Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M. and Paterson, A.H. 2008, Synteny and collinearity in plant genomes, *Science*, **320**, 486–8.
41. Raymond, O., Gouzy, J., Just, J., et al. 2018, The Rosa genome provides new insights into the domestication of modern roses, *Nat. Genet*., **50**, 772–8.
42. Jiang, F., Zhang, J., Wang, S., et al. 2019, The apricot (*Prunus armeniaca* L.) genome elucidates Rosaceae evolution and beta-carotenoid synthesis, *Hortic. Res*., **6**, 128.
43. Zhang, Q.X., Chen, W.B., Sun, L.D., et al. 2012, The genome of *Prunus mume*, *Nat. Commun*., **12**, 3.1318.
44. Wu, J., Wang, Z., Shi, Z., et al. 2013, The genome of the pear (*Pyrus bretschneideri* Rehd). *Genome Res*., **23**, 396–408.
45. Chen, F., Su, L., Hu, S., et al. 2021, A chromosome-level genome assembly of rugged rose (*Rosa rugosa*) provides insights into its evolution, ecology, and floral characteristics, *Hortic. Res*., **8**, 141.
46. Edger, P.P., VanBuren, R., Colle, M., et al. 2018, Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (*Fragaria vesca*) with chromosome-scale contiguity, *Gigascience*, **7**, 1–7.
47. Verde, I., Abbott, A.G., Scalabrin, S., et al. 2013, The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution, *Nat. Genet*., **45**, 487–94.
48. Daccord, N., Celton, J.M., Linsmith, G., et al. 2017, High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development, *Nat. Genet*., **49**, 1099–106.
49. Li, Q., Ai, G., Shen, D., et al. 2019, A *Phytophthora capsici* effector targets ACD11 binding partners that regulate ROS-mediated defense response in Arabidopsis, *Mol. Plant*., **12**, 565–81.
50. Yu, X.F., Han, J.P., Li, L., Zhang, Q., Yang, G.X. and He, G.Y. 2020, Wheat PP2C-a10 regulates seed germination and drought tolerance in transgenic Arabidopsis, *Plant Cell Rep*., **39**, 635–51.
51. Tanaka, Y., Sasaki, N. and Ohmiya, A. 2008, Biosynthesis of plant pigments: anthocyanins, betalains and carotenoids, *Plant J*., **54**, 733–49.
52. Newman, J.D. and Chappell, J. 1999, Isoprenoid biosynthesis in plants: carbon partitioning within the cytoplasmic pathway, *Crit. Rev. Biochem. Mol. Biol*., **34**, 95–106.
53. Rohmer, M. 1999, The discovery of a mevalonate-independent pathway for isoprenoid biosynthesis in bacteria, algae and higher plants, *Nat. Prod. Rep*., **16**, 565–74.
54. Spiller, M., Berger, R.G. and Debener, T. 2010, Genetic dissection of scent metabolic profiles in diploid rose populations, *Theor. Appl. Genet*., **120**, 1461–71.
55. Magnard, J.L., Roccia, A., Caissard, J.C., et al. 2015, Biosynthesis of monoterpene scent compounds in roses, *Science*., **349**, 81–3.
56. Joichi, A., Yomogida, K., Awano, K.I. and Ueda, Y. 2005, Volatile components of tea-scented modern roses and ancient Chinese roses, *Flavour Fragr. J*., **20**, 152–7.
57. Beekwilder, J., Alvarez, H.M., Neef, E., et al. 2004, Functional characterization of enzymes forming volatile esters from strawberry and banana, *Plant Physiol*., **135**, 1865–78.
58. Knudsen, J.T. and Klitgaard, B.B. 1998, Floral scent and pollination in *Browneopsis disepala* (Leguminosae: caesalpinioideae) in western Ecuador, *Brittonia*, **50**, 174–82.