DNA Research, 2018, 25(2), 113–121
doi: 10.1093/dnares/dsx042
Advance Access Publication Date: 16 October 2017
Full Paper

Full Paper

Genome structure of *Rosa multiflora*, a wild ancestor of cultivated roses

Noriko Nakamura[1,†], Hideki Hirakawa[2,†], Shusei Sato[3], Shungo Otagaki[4], Shogo Matsumoto[4], Satoshi Tabata[2], and Yoshikazu Tanaka[1,*]

[1]Suntory Global Innovation Center Ltd, Seika-cho, Soraku-gun, Kyoto 619-0284, Japan, [2]Kazusa DNA Research Institute, Kisarazu, Chiba 292-0818, Japan, [3]Graduate School of Life Sciences, Tohoku University, Aoba-ku, Sendai, Miyagi 980-8577, Japan, and [4]Graduate School of Bioagricultural Sciences, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8601, Japan

*To whom correspondence should be addressed. Tel. +81 50 3182 0451. Fax. +81 774 98 6262.
Email: Yoshikazu_Tanaka@suntory.co.jp

[†]These authors contributed equally to this work.

Edited by Dr. Katsumi Isono

Received 12 June 2017; Editorial decision 15 September 2017; Accepted 19 September 2017

Abstract

The draft genome sequence of a wild rose (*Rosa multiflora* Thunb.) was determined using Illumina MiSeq and HiSeq platforms. The total length of the scaffolds was 739,637,845 bp, consisting of 83,189 scaffolds, which was close to the 711 Mbp length estimated by *k*-mer analysis. N50 length of the scaffolds was 90,830 bp, and extent of the longest was 1,133,259 bp. The average GC content of the scaffolds was 38.9%. After gene prediction, 67,380 candidates exhibiting sequence homology to known genes and domains were extracted, which included complete and partial gene structures. This large number of genes for a diploid plant may reflect heterogeneity of the genome originating from self-incompatibility in *R. multiflora*. According to CEGMA analysis, 91.9% and 98.0% of the core eukaryotic genes were completely and partially conserved in the scaffolds, respectively. Genes presumably involved in flower color, scent and flowering are assigned. The results of this study will serve as a valuable resource for fundamental and applied research in the rose, including breeding and phylogenetic study of cultivated roses.

Key words: *Rosa multiflora*, rose, genome sequencing, gene prediction, flower

1. Introduction

Cultivated roses (*Rosa* × *hybrida*) are by far the most beloved flowers and the most important floricultural crop. The genus *Rosa* contains 120–200 species distributed in the Northern hemisphere. *R. hybrida* constitutes approximately eight *Rosa* species (*R. multiflora*, *R. luciae*, *R. moschata*, *R. damascene*, *R. gallica*, *R. chinensis*, *R. gigantea*, and *R. foetida*)[1–3] by repeated natural and artificial hybridizations. *R. hybrida* is a tetraploid ($2n = 4x = 28$), and the ancestral roses are diploid ($2n = 2x = 14$). As a consequence of natural and artificial hybridization, cultivated roses have various floral characteristics, including intense red, orange, and yellow colors, ever-flowering and larger inflorescences with a large number of petals. Roses have been partly characterized focusing on ornamentally important characters such as flower color, scent, flowering, floral morphogenesis, and development. Roses generally share common mechanisms of these characters with other flowering plants. Roses also contain

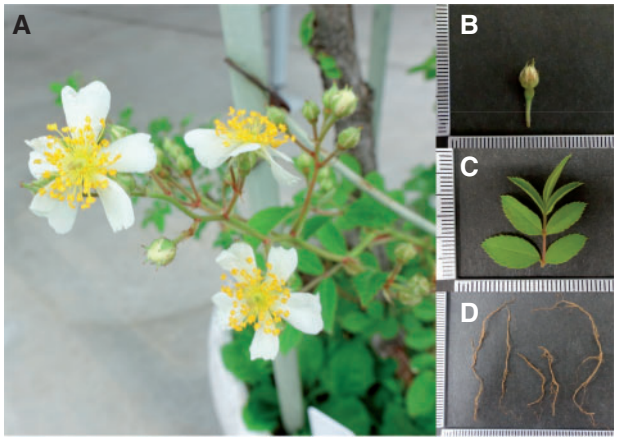

Figure 1. *R. multiflora* used in this study (A). Total RNA was prepared from the petals of buds (B), leaves (C) and roots (D) for RNA-Seq analysis.

unique enzymes such as anthocyanin 5, 3-glucosyltransferase,[4] nucleoside diphosphate linked some moiety X hydrolase 1 (Nudix 1) leading to monoterpenes,[5] and phenylpyruvate decarboxylase (RyPPDC) leading to 2-phenylethanol (2PE).[6]

The *Rosaceae* family contains many important fruit plants and some genome structures have been studied, such as apple (*Malus* × *domestica*),[7] woodland strawberry (*Fragaria vesca*),[8] Japanese apricot (*Prunus mume*),[9] peach (*Prunus persica*),[10] pear (*Pyrus bretschneideri*),[11] and European pear (*Pyrus communis*).[12] Recently, the *Rosa roxburghii* genome, a Chinese medicinal rose, was surveyed.[13] The Rose Genome Sequence Initiative is presently obtaining a high-quality genome sequence of *R. chinensis* cultivar Old Blush and provides RNA-Seq data.[14] The scaffold sequences of *Rosa* × *damascena* (accession number PRJNA322107) and Illumina raw reads from *R. dumalis*, *R. inodora*, and *R. canina* (PRJEB15546) are available from the NCBI SRA database. Currently, the whole genome sequence of an ancestral species of *R. hybrida* has not been revealed.

R. multiflora belonging to the section *Synstylae* is native to eastern Asia, including Japan. It is a thorny perennial shrub and exhibits clusters of white or pale fragrant flowers of five petals. *R. multiflora*, derived from Japan, was utilized to breed modern cultivated roses[15] to confer clustering fluorescence to *R. hybrida*. Its resistance locus (*Rdr1*) to black spot caused by *Diplocarpon rosae* Wolf has been introgressed into *R. hybrida*.[16] A genomic region of 265,477 bp containing *Rdr1* with a cluster of nine highly related TIR-NBS-LRR candidate genes has been reported.[16] The nuclear (2 C) DNA amounts of *R. multiflora* has been estimated to be 1.65 pg,[17] indicating its haploid genome size is approximately 750 Mb.

To deepen fundamental understanding of *R. multiflora* and related species, structural analysis of the whole genome of *R. multiflora* was performed. This genomic study will also be a valuable resource for rose breeding, in combination with the genetic map[18] and pave the way to clarify complex pedigree of the cultivated roses in terms of genome level.

2. Materials and methods

2.1. Materials

R. multiflora studied here was obtained from Keisei Rose Nurseries (Chiba, Japan) (Fig. 1). This plant cultivar originated from Sawara, Chiba prefecture, Japan. The flavonoids of the petals were analysed previously as described.[19] Genomic DNA was prepared from the young leaves using DNeasy Plant Maxi Kit (QIAGEN, Valencia, USA). Total RNA was prepared from the petal of bud, young leaf, and young root of *R. multiflora* (Fig. 1B, C, and D) and from the young petal and young leaf of the *R. hybrida* cultivar 'Rote Rose' using RNeasy Plant Mini Kit (QIAGEN, Valencia, USA).

2.2. Shotgun sequencing

2.2.1. Next-generation sequencing

Shotgun sequencing was carried out using HiSeq 2000 and MiSeq platforms (Illumina Inc., CA, USA). The paired-end (PE) library with insert size of approximately 500 bp was prepared by TruSeq Nano DNA LT Sample Prep Kit. Mate-pair (MP) libraries with insert sizes 2, 5, 10, and 20 kb were constructed by Nextera Mate Pair Sample Prep Kit and GS FLX Titanium Paired End Adaptor Set. These samples were run on HiSeq 2000 and MiSeq with 101 and 301 cycles sequencing kits, respectively. Sequencing of transcripts (RNA-Seq) was also performed for *R. multiflora* and *R. hybrida* cultivar 'Rote Rose', the most common rose cultivar in Japan. The PE libraries of RNA-Seq sampled from bud, leaf, and root were sequenced by MiSeq sequencer.

2.2.2. Quality control

The read quality was checked by FastQC 0.11.2.[20] Nucleotides with quality value <10 and adaptor sequences at 3′ termini of reads were trimmed by PRINSEQ 0.20.4[21] and FASTX-toolkit 0.0.14 (http://hannonlab.cshl.edu/fastx_toolkit), respectively. After trimming, PE and MP reads with >100 bases and PE reads with size of 250–300 bases were selected for sequencing by HiSeq 2000 and MiSeq, respectively. Reads shorter than 99 bases and including unknown nucleotides were excluded.

2.2.3. Estimation of genome size

The genome size of *R. multiflora* was estimated using HiSeq 2000 and MiSeq PE reads with *k*-mer size = 17. The *k*-mer distribution was investigated using Jellyfish 2.1.3.[22] Genome size was estimated using the peak in the *k*-mer frequency distribution curve according to the method used in a previous study.[23]

2.3. Assembly, expression level analysis, and SNP detection

The trimmed PE and MP reads were used for genome assembly by SOAPdenovo rev240 (-M 1)[24] with *k*-mer sizes, 71, 81, and 91. After assembly, gaps on scaffolds were closed by GapCloser 1.10 (http://soap.genomics.org.cn/soapdenovo.html) ($p = 31$). The scaffolds were subjected to BLASTN searches with *E*-value cutoff of 1*E*-10 and length coverage $\geq$10% against bacteria, fungi, and human genome sequences (hg19) from NCBI, vector sequences from UniVec (http://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/), genome sequences of chloroplast of *Fragaria vesca* (NC_015206.1) and mitochondria of *Arabidopsis thaliana* (NC_001284.2), and PhiX sequence used in Illumina sequencing. Scaffolds exhibiting homology to these were excluded as contamination.

RNA-Seq reads sampled from bud, leaf, and root of *R. multiflora* were assembled by Trinity r20140717.[25] At same gene locus, several contigs derived from splicing variants were predicted; therefore, the contig with the highest IsoPct value calculated by RSEM 1.2.15[26] was selected as the transcript. The contigs thus obtained were mapped onto the scaffolds by BLAT v34[27] with $\geq$90% similarity and score $\geq$200 (–maxIntron = 10,000). According to the mapping results, scaffolds were connected by L_RNA_scaffolder.[28] As a result, scaffolds longer than 300 bases were selected and designated

RMU_r2.0. Repetitive sequences were detected by RepeatScout 1.0.5[29] and RepeatMasker 4.0.3 (http://www.repeatmasker.org) according to the method used in a previous study.[23]

Expression levels of the genes in bud, leaf, and root were investigated for *R. multiflora*. The RNA-Seq reads were mapped onto the scaffolds of RMU_r2.0 by TopHat v2.0.14.[30] The FPKM (fragments per kilobase of exon per million fragments mapped reads) value of the genes were calculated by Cufflinks v2.2.1.[31] RNA-Seq reads of *R. hybrida* cultivar 'Rote Rose' were used for detection of SNPs distinguishing *R. multiflora*. The RNA-Seq reads were mapped onto the scaffolds of RMU_r2.0 by TopHat v2.0.14.[30] The BAM files obtained were used for SNP detection by SAMtools v0.1.19.[32]

2.4. Authenticity

The full BAC sequence of the black spot resistance *muRdr1* gene locus of *R. multiflora* line 88/124-46[16] was obtained from NCBI nuccore database (HQ455834.1; 265,477 bp). The nine TIR-NBS-LRR resistance proteins, muRdr1A-muRdr1I, were encoded in the BAC sequence. The genes were mapped onto the scaffolds of RMU_r2.0 by BLAT with ≥95% similarity and score ≥200 (–minIdentity = 95). To compare the genic regions among the related species, the EST sequences of *R. hybrida* (12,649 sequences; Supplementary Table S1), *R. luciae* (1,936 sequences; Supplementary Table S1), and *R. virginiana* (5,978 sequences; Supplementary Table S1) obtained from NCBI's dbEST were mapped onto the genome sequence of RMU_r2.0 by BLAT with ≥95% similarity and score ≥100. Conservation of the core eukaryotic genes and single-copy orthologous genes were investigated using CEGMA v2.5[33] and BUSCO ver. 1.1b,[34] respectively. In CEGMA, genome completeness was estimated by using 248 CEGs (Core Eukaryotic Genes) to classify them into complete and partial genes. In BUSCO, genome completeness was estimated by using single-copy orthologous genes selected from OrthoDB to classify them into complete genes (single-copy and duplicated), fragmented genes, and missing genes.

2.5. Gene assignment and annotation

The RNA-Seq reads were mapped onto the scaffolds of RMU_r2.0 with TopHat 2.0.14[30] to generate a BAM file. This file was used for making a training set by BRAKER1 v1.3.[35] GeneMark-ET 4.21,[36] and Augustus 3.0.3[37] were initially applied to build the training set, and Augustus 3.0.3 was applied to predict genes using the training set. The predicted genes were subjected to homology searches against NCBI NR database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz) and translated genes of *A. thaliana* in TAIR10 (https://www.arabidopsis.org) using BLASTP with an *E*-value cutoff of 1*E*-10. Domain searches against InterPro (http://www.ebi.ac.uk/interpro/) were conducted using InterProScan[38] with an *E*-value cutoff of 1.0. Genes related to transposable elements (TEs) were inferred according to the results of BLAST searches against the NCBI NR database, and domain searches against InterPro and GyDB 2.0[39] using hmmsearch in HMMER 3.0[40] with an *E*-value cutoff of 1.0. Transfer RNA genes (tRNAs) were predicted by tRNAscan-SE 1.23.[41] Ribosomal RNA genes (rRNAs) were predicted by HMMER 3.0 searches against the Rfam 11.0 database.

The predicted peptide sequences of *R. multiflora* were searched to assign enzymes/proteins to various biosynthetic pathways by BLASTP using counterpart protein sequences as queries, and one or a few best-matched sequences with low *E*-values were selected. In the case of phylogenetic analysis of MADS-box genes, expansins, xyloglucan endotransglycosylase/hydrolases (XTHs) or aquaporins, we used BLASTP and keyword search (the word 'MADS', 'expansin', 'xyloglucan endotransglucosylase/hydrolase', or 'aquaporin' was used) in the *Rosa multiflora* Genome DataBase (http://rosa.kazusa.or.jp). For BLASTP search, we used amino acid sequences of each genes characterized in model plants as queries with *E*-value cutoff of 1E-20. Multiple sequence alignments were constructed by DDBJ CLUSTALW version 2.1 (http://clustalw.ddbj.nig.ac.jp/) with default parameters. Phylogenetic trees were generated using the neighbor-joining method included in the DDBJ CLUSTALW version 2.1 with default parameters and circular cladograms were constructed using Dendroscope 3.[42]

2.6. Functional analyses of the predicted genes

The predicted genes were classified into the functional categories defined in NCBI's 'euKaryotic clusters of Orthologous Groups (KOG)'[43] by BLASTP searches with an *E*-value cutoff of 1*E*-4. The genes were mapped onto the KEGG reference pathways by BLAST searches against the KEGG GENES database (http://www.genome.jp/kegg/genes.html) and UniProt (TrEMBL + Swiss-Prot) database (http://www.uniprot.org) with an *E*-value cutoff of 1*E*-80, length coverage of 25%, and identity of 50%. The genes were also mapped onto the KEGG reference pathways of *F. vesca* (v2.0a1), *P. persica* (peach; v2.0a1), and *Malus* × *domestica* (apple; v1.0p). The genes of the four plant species were compared by OrthoMCL v2.0.9.[44] The genome sequence, CDS and pep sequences, and annotation of *F. vesca* (v2.0a1), peach (v2.0a1), and apple (v1.0p) were obtained from the Genome Database for Rosaceae (GDR; https://www.rosaceae.org). The Gene Ontology (GO) categories were assigned to the genes based on the raw files obtained by InterProScan. The distribution of the genes in GO categories was investigated according to GOslim (http://www.geneontology.org/page/go-slim-and-subset-guide).

3. Results and discussion

3.1. Sequencing of the rose genome

3.1.1. Shotgun sequencing of the rose genome

The 1.1 G PE reads were obtained by HiSeq 2000 with length 100 bp and 178 M PE reads for MiSeq with length 301 bp. The MP reads of insert sizes, 2, 5, 10, and 20 kb, were obtained by HiSeq 2000 for 501, 610, 425, and 394 M reads with length 101 bp. The obtained reads are summarized in Supplementary Table S2. The quality of reads was checked using FastQC, and quality trimming and adaptor trimming were performed by PRINSEQ and FastX-toolkit, respectively. The one nucleotide on 3′ termini was trimmed because of low quality. Finally, reads with lengths 100 and 250–300 bp were selected for HiSeq 2000 and MiSeq reads, respectively, and divided into paired and single reads. After quality control, 95.3% of MiSeq PE reads and 88.7% of HiSeq 2000 PE reads were used for assembly. For MP reads with insert sizes, 2, 5, 10, and 20 kb obtained by HiSeq 2000, 26.3%, 4.4%, 9.3%, and 4.9% were used for assembly, respectively.

3.1.2. Assembly authenticity and sequence features

The *k*-mer frequency distribution curve (*k*-mer = 17) derived from PE reads is shown in Supplementary Fig. S1. As a result, the genome size was estimated at 1,087,968,027 and 711,129,940 bases using the two peaks at multiplicity = 117 (coverage = 133.7) and 179 (coverage = 204.5), respectively. According to the peak (multiplicity = 179), the haploid genome size was estimated to 711 Mb, which was close to the estimated size,[17,23] and thus the peak at multiplicity = 117 may reflect the presence of heterozygosity in *R. multiflora*.

Table 1. Genomic feature of RMU_r2.0 and RMU_r2.0_cds

	RMU_r2.0 (genome)	RMU_r2.0_cds (CDS)
Number of sequences	83,189	67,380
Total length (bases)	739,637,845	66,058,172
Average length (bases)	8,891	980
Max length (bases)	1,133,259	20,538
Min length (bases)	501	149
N50 length (bases)	90,830	1,272
G+C%	38.9	45.9
Repeat%	56.4	–
Complete genes	–	54,893
Partial genes	–	12,487

The numbers of raw and trimmed reads are summarized in Supplementary Table S2. The trimmed reads were applied to the assembly using SOAPdenovo2 with *k*-mer sizes = 71, 81, and 91. From the results of the assemblies, the N50 lengths using *k*-mer sizes 71, 81, and 91 were 30,732, 35,285, and 28,059 bases, respectively. Judging from the N50 lengths, the scaffolds assembled with *k*-mer size = 81 were used for further analysis (Supplementary Table S3). The gaps on the scaffolds were closed by GapCloser 1.10. The RNA-Seq from bud, leaf, and root was assembled by Trinity, and splicing variants were excluded by RSEM (Supplementary Table S4). The contigs were used for scaffolding using L_RNA_scaffolder. As a result, 158,733 scaffolds with total length 767,886,425 and N50 length 86,097 bp were obtained (Supplementary Table S3). Finally, 75,439 scaffolds shorter than 500 bp and 105 scaffolds with probable contamination hit against chloroplast, mitochondrial, fungal and bacterial genomes were excluded, and the resultant 83,189 scaffolds were determined as representing the draft genome sequence, RMU_r2.0 (Table 1). The total length of RMU_r2.0 was 739,637,845 bp, and N50 length was 90,830 bp. The scaffolds were prefixed with 'Rmu' followed by a seven-digit identifier and sequence version (e.g. Rmu_ssc0000001.1). Considering the genome size estimated by distribution of *k*-mer frequency, the total length of the assembled genome sequence was somewhat longer, probably due to heterozygosity.

Authenticity of the assembled genome sequence was performed by similarity searches for the full BAC sequence of *R. multiflora* and for EST sequences of *R. hybrida*, *R. lucieae*, and *R. virginiana*. In the BAC sequence, only nine TIR-NBS-LRR resistance genes were annotated. Therefore, gene prediction was conducted to the BAC sequence by the same method applied to the scaffolds. As a result, 53 genes were predicted from the BAC sequence, and 10 genes of them were homologous to TIR-NBS-LRR resistance genes. Synteny between Rmu_sc0000110.1 and the BAC sequence was investigated using nucmer,[45] as shown in Supplementary Fig. S2. There are two regions without similarity between them. The red and blue bars indicate the genic regions on plus and minus strands, respectively. The gray bar indicate the unknown nucleotides (Ns). The region from 160 kb to 200 kb on the BAC sequence corresponded to the region with unknown nucleotides on Rmu_sc0000110.1. On the other hand, the region from 20 kb to 120 kb was not similar, which encoded TIR-NBR-LRR resistance genes. The BAC clone encodes TIR-NBS-LRR resistance genes, and the genome region of TIR-NBS-LRR resistance genes tend to be rearranged.[46] As the strain of the BAC clone (breeding line 88/124-46) is different from our target, the non-conserved region of the BAC clone could be caused by genome rearrangement. However, the half region on 3' terminal without TIR-NBS-LRR genes on the BAC sequence was conserved well. Moreover, 12,003 of 12,649 (94.8%) of EST sequences of *R. hybrida*, 1,455 of 1,936 (75.1%) of EST sequences from *R. lucieae*, and 5,040 of 5,978 (84.3%) of EST sequences in *R. virginiana* were mapped onto RMU_r2.0, respectively.

Authenticity of the assembled genome sequence was also verified by use of CEGMA[33] and BUSCO[34] programs. The results of CEGMA and BUSCO were shown in Supplementary Table S5. Of the 248 core genes, 228 (91.9%) were completely conserved, while 15 (6.1%) were partially conserved, according to the CEGMA output. In contrast, 845 of the 956 genes defined in BUSCO program (88.4%) were classified as 'complete single-copy.' 348 genes (36.4%) were categorized as 'complete duplicated,' 40 (4.2%) were classified as 'fragmented,' and 71 genes (7.4%) were classified as 'missing' when analysed by BUSCO. The 91.9% and 88.4% of the genes defined by CEGMA and BUSCO were classified into complete structure, which indicate that the genes were conserved with high rates judged from the status of the genome assembly. In BUSCO analysis, 348 (36.4%) complete duplicated genes were detected. This high rate might be due to the high heterozygosity in RMU_r2.0. These results indicated that the core genes and single-copy orthologous genes might be conserved in RMU_r2.0. The core genes duplicated in RMU_r2.0 might be derived from the contigs separated by heterozygosity.

3.2. Characteristic features of the rose genome

3.2.1. Repetitive sequences

The known and unique repetitive sequences identified in RMU_r2.0 are summarized in Supplementary Table S6. Total length of the known repeats was 417,242,576 bp (56.4% of the total) and in which Class I LTR elements were frequently present. In contrast, the unique repeats that have not been sequenced were newly identified in our analysis; the total length of these was 290,260,400 bp (39.2% of the total).

3.2.2. Prediction of protein-encoding genes and annotation

According to the mapping results for RNA-Seq reads, 3,622,550 of 5,123,157 read pairs (70.7%) were mapped onto the draft genome sequence (RMU_r2.0). The training set for *R. multiflora* was constructed by BRAKER1,[35] and used for gene prediction by Augustus 3.0.3.[47] As a result, we predicted 178,512 genes on the genome sequence. According to the BLAST search against the NCBI NR database and by domain searches against InterPro, 67, 380 genes were selected with support by sequence alignment evidence. The first gene set was called after RMU_r2.0.cds, and second one was called after RMU_r2.0.braker1.cds. The statistics of the predicted genes are summarized in Supplementary Table S7. The total length of the CDSs was 66,058,172 bp with 45.9% GC content. This number of CDSs was higher compared with other *Rosaceae* plants, which may be caused by the presence of genes derived from duplicated contigs due to heterozygous genome regions that self incompatibility of *R. multiflora* results in. Kajitani et al. simulated the genome assembly of *Caenorhabditis elegans* with various levels of heterozygosity in Illumina reads ranged from 0.1 to 2.0%, and they indicated that the lengths of the contigs and scaffolds were shorter as higher heterozygosity.[48] This means that the genes would be partial in the case of high heterozygosity. The gene name

was prefixed with a seven-digit identifier followed by scaffold or contig number, for example, Rmu_ssc0000001.1_g000001.1. According to the presence of start or stop codons, genes were tagged as partial (with start or stop codons or without start or stop codons) or pseudogenes (presence of stop codons in CDS). Genes less than 50 amino acids were tagged as short. Genes having similarity to NR database entries by BLAST with *E*-values ≤1*E*-20 and identity ≥70% were tagged with 'f,' and those with *E*-values ≤1*E*-20 and identity <70% were tagged with 'p.' Genes having hits against InterPro with *E*-values ≤1.0 were labeled with 'd.' In RMU_r2.0.braker1.cds, genes having similarity to transposable elements were tagged with 'TE.' The number of genes with TE tag was 46,505. In RMU_r2.0.cds, the numbers of the tags '/fd (tags f and d),' '/pd (tags p and d),' '/f-,' '/p-,' and '/-d' were 24,121, 16,461, 2,416, 5, 259, and 6,636. The numbers of the tags for partial genes '/partial/fd,' '/partial/pd,' '/partial/f,' '/partial/p,' '/partial/d' were 6,842, 2,067, 1,240, 900, 1,438, respectively. The numbers of 5.8 S, 18 S, and 25 S rRNA genes were 4, 22, and 31, respectively. The tRNA and rRNA genes predicted in *R. multiflora*, *F. vesca*, and *P. persica* are compared in Supplementary Tables S8 and S9, respectively. The distributions of KOG functional categories of *R. multiflora*, *F. vesca*, and *P. persica* were similar (Supplementary Fig. S3). The numbers of the genes mapped onto KEGG metabolic pathways classified into '1. Metabolism' are shown in Supplementary Table S10. The 17,677 genes (26.2%) of *R. multiflora* were mapped onto 346 of the 476 metabolic pathways in the KEGG database, whereas the 8,262 (24.6%), 11,710 (39.3%), 12,753 (48.2%), and 12,934 (46.7%) genes of *F. vesca*, *M.* × *domestica*, *P. persica*, and *A. thaliana* were mapped onto 344, 342, 344, and 345 pathways, respectively. The pathways were categorized as to which genes in the *R. multiflora* genome were uniquely mapped and were as follows: 'Ascorbate and aldarate metabolism' in '1.1. Carbohydrate metabolism,' 'Methane metabolism' in '1.2 Energy metabolism,' 'Riboflavin metabolism' in '1.8 Metabolism of cofactors and vitamins,' 'Monoterpenoid biosynthesis' in '1.9 Metabolism of terpenoids and polyketides,' 'Isoquinoline alkaloid biosynthesis' in '1.10 Biosynthesis of other secondary metabolites.'

3.2.3. Flavonoid biosynthetic genes

Pink- to red-flower colors of roses are derived from cyanidin or pelargonidin glucosides belonging to anthocyanins, a class of colored flavonoids (Supplementary Fig. S4). Rose flowers do not contain the delphinidin or flavone that is common in blue or violet flowers. Anthocyanin biosynthesis leading to anthocyanidin 3-glucoside is well conserved in higher plants.[49] The *R. multiflora* genome contains amino acid sequences exhibiting high identity to reported biosynthetic enzymes (Supplementary Fig. S4).

R. multiflora petals contain large amounts of kaempferol (3.975 ± 0.183 mg/g fresh petals), small quantities of quercetin (0.109 ± 0.014), and cyanidin (0.001 ± 0.001). Flavones and 3′, 5′-hydroxylated flavonoids such as delphinidin and myricetin were not detected. The flavonoid profiles indicate the presence of flavonol synthase (FLS) and flavonoid 3′-hydroxylase (F3′H) (Supplementary Fig. S4) and the absence of flavonoid 3',5'-hydroxylase (F3′5′H) and flavone synthase (FNS) in *R. multiflora*. Genomic genes corresponding to FLS and F3′H were identified in the genome (Supplementary Fig. S4), and those corresponding to F3′5′H and FNS were not found. The genes involved in regulation of flavonoid biosynthesis and vacuolar transport will be reported separately. The flavonoid biosynthetic pathway includes plural cytochromes P450 (P450) and UDP-sugar dependent glucosyltrasnferases/glycosyltransferases (GT). The *R. multiflora* genome contains 677 P450 and 507 GT ORFs in the scaffold sequences of RMU_r2.0. These numbers exceed those in other plant genomes, confirming that the *R. multiflora* genome is heterogeneous. Many GT and P450 genes are found to form clusters in the same scaffolds. Rmu_sc0005080.1 contains 11 GT genes and Rmu_sc0000698.1 contains 10 P450 genes (Supplementary Tables S11 and S12). Such clusters of P450 and GT genes are revealed in *Glycyrrhiza uralensis*.[50] P450 and GT co-localize in eight scaffolds in *G. uralensis* and in four scaffolds of the *R. multiflora* genome (Supplementary Table S13).

3.2.4. Floral scent and carotenoid-related genes

Rose floral scent compounds are mainly benzenoids such as 2-phenylethanol (2PE) and terpenoids, including geraniol. Scent compounds of modern roses are derived from their ancestral wild roses. 2PE is synthesized by two pathways: one is via aromatic amino acid decarboxylase (AADC)[51] and phenylacetaldehyde reductase (PAR) in winter[52] and the other is via phenylpyruvate decarboxylase (PPDC) in summer.[6] The biosynthetic pathway of floral scent compounds, the relevant enzymes, and corresponding *R. multiflora* genes are summarized in Supplementary Fig. S5. *R. multiflora* predominantly produces 2PE but not 3, 5-dimethoxytoluene (DMT) or 1, 3, 5-trimethoxybenzene (TMB).[53] 2PE has a rose-like floral note and is one of the key scent compounds in roses.[51] Genes corresponding to AADC, PAR, and PPDC are found in the genome (Supplementary Fig. S5). Eugenol synthase and eugenol methyltransferase genes were also found.

The 'tea scent' compound of tea roses has been shown to be DMT, which is derived from *R. chinensis* and synthesized by catalysis of two closely related orcinol *O*-methyltransferases.[54] Phenolic methyl ethers such as DMT or TMB, the characteristic 'tea scent' compounds of tea roses, have an earthy and spicy note.[54] Only one orcinol methyltransferase gene (Rmu_sc0002707.1_g000004.1) was found in the *R. multiflora* genome, while Chinese roses producing DMT and TMB contain two closely related orcinol methyltransferase (*OOMT1* and *OOMT2*) genes.[54] This is consistent with the absence of DMT or TMB in *R. multiflora*. It has been suggested that *R. chinensis* OOMT1 contains a tyrosine residue at amino acid 127, whereas OOMT2 has a phenylalanine residue at this position.[54] It has been suggested that OOMT1 does not catalyze 3-methoxy-5-hydroxytoluene (Supplementary Fig. S5) due to structural hindrance by the hydroxyl group of the tyrosine residue but OOMT2 does.[54] The gene Rmu_sc0002707.1_g000004.1 encoding phenylalanine at this position may be indicative of the presence of *OOMT2* catalyzing methylation of 3-methoxy-5-hydroxytoluene. Terpenoids are the largest floral scent group and are synthesized from prenyl diphosphate precursors by terpene synthases. In particular, monoterpenes (C_{10}) such as geraniol or linalool, which are synthesized from geranyl diphosphate (GPP) and have flower-like notes, are major scent components in some modern roses.[55] Although several genes encoding linalool, nerolidol, or α-pinene synthase homologues were found here, these compounds are not known to produce in *R. multiflora* flowers. It is interesting that *R. multiflora* contains two genomic genes encoding nucleoside diphosphate linked some moiety X hydrolase 1 (NUDX1) which is suggested to be involved in an alternative pathway[5] to synthesize geraniol in spite of the absence of geraniol and its derivatives in *R. multiflora*. Carotenoid cleavage dioxygenase 1 (CCD1) cleaves β-carotene at the 9–10 and the 9′–10′ positions and generates two β-ionones (C_{13} product), which has violet-like notes.[56] The *CCD1* gene leading to β-ionone was also assigned (Supplementary Fig. S5).

3.2.5. MADS-box genes related to ABCDE model for floral development

The ABCDE model has been developed for identification of different floral organs, namely sepals, petals, stamens, and carpels, and these organs are categorized as so-called A-class for sepal and petal specification, B-class for petal and stamen specification, C-class for stamen and carpel specification, D-class for carpel and ovule specification, and E-class for sepal, petal, stamen, and carpel specification by homeotic genes (mostly MADS-box-genes).[57] Until now, 11 MADS-box-genes have been identified from wild and cultivated roses: class A genes *RhAP1-1, RhAP1-2*, and *RhFUL*[58]; class B genes *MASAKO BP*, *MASAKO B3*, and *MASAKO euB3*[59,60]; class C/D genes *MASAKO C1* and *MASAKO D1*[61]; and class E genes *RhSEP3*, *MASAKO S1*, and *MASAKO S3*.[62,63] Attenuated *MASAKO C1* (*RhAG*) expression under low temperature condition causes an additional petal and petaloid stamens formation in cultivated rose,[64] and severe reduction of *MASAKO C1* expression was observed sterilized (anemone type) *R. luciae* flowers by our analysis (Supplementary Fig. S6). However, the *R. multiflora MASAKO C1* (Rmu_sc0003469.1_g000007.1) have an intact open reading frame without a frame shift or transposon insertion. This is consistent with its normal flower phenotype.

R. multiflora contains 94 MADS-box genes, including three class A, four class B, two class C/D, one class D, five class E, and three *SHORT VEGETATIVE PHASE* (*SVP*) genes (Supplementary Table S14). The 15 class ABCDE genes belonged to two clades consisting of class B, and class A, C/D and E except for one class D (Rmu_sc0000512.1_g000020.1) and one class E (*SEP4*) (Rmu_sc0003558.1_g000005.1) (Supplementary Fig. S7).

3.2.6. Everblooming character

The everblooming characteristic is one of the most important in modern roses and originated from everblooming sport of *R. chinensis*. A retrotransposon insertion in the *KSN* (terminal flower 1 homologue) gene resulted in this characteristic.[65] Although Rmu_sc0010986.1_g000002.1 protein exhibited two amino acid residue differences from *R. chinensis* KSN protein and the *R. multiflora KSN* gene has a shorter first intron and a longer third intron than the *R. chinensis KSN* gene (Supplementary Fig. S8), no retrotransposon insertion is found in the Rmu_sc0010986.1_g000002.1 sequence. This is consistent with once flowering phenotype of *R. multiflora*.

3.2.7. Genes related to flower opening of rose petals

Fully extended petals and a long vase life are prerequisites for increasing the ornamental value of rose flowers. For extending petals, expansion of the petal cells plays a pivotal role. Expansins, XTHs and aquaporins participate in this process by loosening the cell wall or mediating influx of water into cells.[66] Studies in Arabidopsis and other model plants disclose that these three proteins comprise a multigene superfamily. In rose, three aquaporin, four expansin and four XTHs genes have been identified as relevant to the expansion of petal cells.[67–71]

BLAST and keyword search identified 65 aquaporins, 47 expansins and 60 XTHs (Supplementary Tables S15–17). In aquaporins, 55 out of 65 genes were predicted to have at least two transmembrane helix, which is conserved in aquaporin homologues (Supplementary Table S15). Through the phylogenetic analysis aquaporins excluding seven genes annotated with partial by Augustus 3.0.3 classified into 17 plasma membrane intrinsic proteins, 11 tonoplast intrinsic proteins, 10 nodulin 26-like intrinsic proteins, 6 small basic intrinsic proteins (SIPs) and 4 X intrinsic proteins (Supplementary Fig. S12 and S15). Similarly, expansins annotated with InterProScan accession PR01225 or PR01226 were classified into 3 subfamilies as expected (Supplementary Fig. S9 and Supplementary Table S16). RNA-Seq analysis predicted that 21 aquaporins, 13 expansins and 22 XTHs were expressed in the flower bud, among which 10 genes show particularly high expression level (RPKM > 50) (Supplementary Figs S9–S12 and Supplementary Tables S15–S17). Because seven highly expressed genes were not clustered with known rose genes, it may be interesting to investigate whether these genes actually act as a novel regulator for the expansion of petal cells in rose in the future analysis.

3.3. Comparative genomics

3.3.1. Comparison with other rosaceae genomes at the gene level

The genes predicted in *R. multiflora*, *F. vesca*, *P. persica*, *M.* × *domestica* were compared by a clustering method using OrthoMCL.[44] In *R. multiflora*, the 67,380 genes were classified into intrinsic and partial genes and were used for this comparison. The number of clusters is shown in Supplementary Fig. S13. In the common region, 7,665 clusters were included. In the clusters, 18,956, 10,877, 13,187, and 14,069 genes were included for *R. multiflora*, *F. vesca*, *P. persica*, and *M.* × *domestica*, respectively. The number of clusters shared in common in *R. multiflora* and *F. vesca*, *R. multiflora* and *P. persica*, and *R. multiflora* and *M.* × *domestica* were 1,287, 904, and 241, respectively. This indicates that *R. multiflora* is closely related to *F. vesca*, *P. persica*, and *M.* × *domestica* in a stepwise manner. The distances corresponded well to the phylogenetic relationship in Rosaceae reported by Xiang et al.[72] The number of clusters uniquely found in *R. multiflora* were 2.5 times (3,482 in *R. multiflora*/1,397 in *F. vesca*) higher than that in *F. vesca*. This was due to the heterozygosity in *R. multiflora*. In addition, the number of genes in the clusters uniquely found in *R. multiflora* was 3.3 times (14,663 in *R. multiflora*/4,482 in *F. vesca*) higher than that in *F. vesca*, which means that the duplicated or partial genes were included in *R. multiflora* more than in *F. vesca*.

3.3.2. Comparison with the apple genome at the pseudomolecule level

To investigate possible syntenic relationships among *R. multiflora* and other rosaceous taxa genomes, the status of conservation of relative gene positions was surveyed using the scaffolds of rose genomic sequences. Among the 3,932 scaffolds with five or more predicted genes, conservation of the relative positions of three or more genes was observed in 1,968 scaffolds (50%) and 2,312 scaffolds (59%) against genes predicted in the *F. vesca* and *P. persica* genomic sequences, respectively (Supplementary Table S18). It appears that a significant degree of micro-synteny can be expected within the family Rosaceae.

By anchoring scaffolds of rose genome with SSR marker information in the recent integrated genetic map of roses,[73–75] macro-syntenic relationships between *R. multiflora* and other *Rosaceae* genomes were investigated. A total of 160 scaffolds, with 17.9 Mb total length, were anchored on the seven linkage groups of *R. multiflora*. Among the 127 markers with corresponding scaffolds, 33 markers (26%) correspond to two scaffolds, which presumably indicated the presence of redundant scaffolds corresponding to heterozygous alleles (Supplementary Table S19). The micro-syntenic relations identified on the anchored scaffolds revealed macro-level synteny between

R. multiflora and *F. vesca*, with entire chromosome level synteny between *R. multiflora* linkage group (RG) 1 and *F. vesca* chromosome (FC)7, RG4 and FC4, RG5 and FC3, RG6 and FC2, and RG7 and FC5. A single large translocation was evident, as scaffolds anchored on RG2 showed syntenic relation to either FC1 or the top portion of FC6 and those on RG3 showed syntenic relation to bottom portion of FC6. Macro-syntenic relationships identified between *R. multiflora* and *P. persica* corresponded well to the syntenic relationship between *F. vesca* and *P. persica*,[8] that is, RG1 and the *P. persica* linkage group (PG)2, RG2-PG3/PG7, RG3-PG6, RG4-PG1, PG5-PG4/PG6, PG6-PG1/PG8, and PG7PG1/PG5.

3.3.3. SNP analysis of wild and cultivated roses

To examine molecular similarities among wild rose, *R. multiflora*, and cultivated rose, *R. hybrida*, transcriptome reads of *R. hybrida* cultivar 'Rote Rose' were mapped to our *R. multiflora* genome sequence. A total of 55,086 *R. multiflora* genes were mapped by 20 or more *R. hybrida* transcriptome reads, and 198,807 SNPs were identified on these mapped regions, corresponding to 201 Mb (~27%) of the genome. Although the average SNP density could be calculated as one in every 1.01 kb, a higher SNP density might be expected across the entire genome if nontranscribed regions were also included in this analysis.

3.4. Databases

The genomic and gene sequences and their annotation are available at '*Rosa multiflora* Genome DataBase (http://rosa.kazusa.or.jp).' Users can input query sequences to perform BLAST searches against the genomic and gene sequences (transcripts, CDSs, and proteins) on the 'BLAST' page. Keyword search of the BLAST search result against NR and TAIR10 pep for each gene is available on the 'KEYWORD' page. At the 'DOWNLOAD' page, data for the genomic and gene (cds, pep, and transcripts) sequences, annotation file (gff3 format), and the InterProScan search results (raw format) can be downloaded.

Availability of data

The Bioproject identifier for *R. multiflora* is PRJDB4738. The accession numbers of the assembled genome sequences of RMU_r2.0 are BDJD01000001-BDJD01083189 (83,189 entries). The SRA accession numbers for the Illumina reads (HiSeq and MiSeq) used in this study are summarized in Supplementary Table S2.

Acknowledgements

The authors are grateful to Professor Ueda of Gifu International Academy of Horticulture for his valuable suggestions about the lineage of *R. hybrida* and to Mr. Takeuchi of Keisei Rose Nurseries for providing the *R. multiflora* plant. The authors are indebted to Dr. Katsumoto and Ms. Nakajima for their kind encouragement during this study. Mr. Ito and Mses. Nakata, Yamauchi, Morimoto, and Akahoshi are acknowledged for their technical assistance.

Conflict of interest

None declared.

Funding

This work was supported by the Kazusa DNA Research Institute Foundation.

Supplementary data

Supplementary data are available at *DNARES* online.

References

1. Wylie, A.P. 1954, The history of garden roses. Part I, *J. Roy. Hort. Soc.*, **79**, 555–71.
2. Wylie, A.P. 1954, The history of garden roses. Part II, *J. Roy. Hort. Soc.*, **80**, 8–24.
3. Wylie, A.P. 1954, The history of garden roses. Part III, *J. Roy. Hort. Soc.*, **80**, 77–87.
4. Ogata, J., Kanno, Y., Itoh, Y., Tsugawa, H. and Suzuki, M. 2005, Plant biochemistry: anthocyanin biosynthesis in roses, *Nature*, **435**, 757–8.
5. Magnard, J.L., Roccia, A., Caissard, J.C., PLANT VOLATILES., et al. 2015, Biosynthesis of monoterpene scent compounds in roses, *Science*, **349**, 81–3.
6. Hirata, H., Ohnishi, T., Tomida, K., et al. 2016, Seasonal induction of alternative principal pathway for rose flower scent, *Sci. Rep.*, **6**, 20234.
7. Velasco, R., Zharkikh, A., Affourtit, J., et al. 2010, The genome of the domesticated apple (*Malus* × *domestica* Borkh.), *Nat. Genet.*, **42**, 833–9.
8. Shulaev, V., Sargent, D.J., Crowhurst, R.N., et al. 2011, The genome of woodland strawberry (*Fragaria vesca*), *Nat. Genet.*, **43**, 109–16.
9. Zhang, Q., Chen, W., Sun, L., et al. 2012, The genome of *Prunus mume*, *Nat. Commun.*, **3**, 1318.
10. 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.
11. Wu, J., Wang, Z., Shi, Z., et al. 2013, The genome of the pear (*Pyrus bretschneideri* Rehd.), *Genome Res.*, **23**, 396–408.
12. Chagne, D., Crowhurst, R.N., Pindo, M., et al. 2014, The draft genome sequence of European pear (*Pyrus communis* L. 'Bartlett'), *PLoS One.*, **9**, e92644.
13. Lu, M., An, H., Li, L. and Jain, M. 2016, Genome survey sequencing for the characterization of the genetic background of *Rosa roxburghii* tratt and leaf ascorbate metabolism genes, *PLoS One*, **11**, e0147530.
14. Foucher, F., Hibrand-Saint Oyant, L. and Hamama, L. 2015, Towards the rose genome sequence and its use in research and breeding, *Acta Hortic*, **1064**, 167–75.
15. Hurst, C.C. 1941, Notes on the origin and evolution of our garden roses, *J. Roy. Hort. Soc.*, **66**, 73–82.
16. Terefe-Ayana, D., Yasmin, A., Le, T. L., et al. 2011, Mining disease-resistance genes in roses: functional and molecular characterization of the *Rdr1* locus, *Front. Plant Sci.*, **2**, 35.
17. Dickson, E.E., Arumuganathan, K., Kresovich, S. and Doyle, J.J. 1992, Nuclear DNA content variation within the Rosaceae, *Am. J. Bot.*, **79**, 1081–6.
18. Vukosavljev, M., Arens, P., Voorrips, R.E., et al. 2016, High-density SNP-based genetic maps for the parents of an outcrossed and a selfed tetraploid garden rose cross, inferred from admixed progeny using the 68k rose SNP array, *Hortic. Res.*, **3**, 16052.
19. Katsumoto, Y., Fukuchi-Mizutani, M., Fukui, Y., et al. 2007, Engineering of the rose flavonoid biosynthetic pathway successfully generated blue-hued flowers accumulating delphinidin, *Plant Cell Physiol.*, **48**, 1589–600.
20. Leggett, R.M., Ramirez-Gonzalez, R.H., Clavijo, B.J., Waite, D. and Davey, R.P. 2013, Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics, *Front. Genet.*, **4**, 288.
21. Schmieder, R. and Edwards, R. 2011, Quality control and preprocessing of metagenomic datasets, *Bioinformatics*, **27**, 863–4.
22. Marçais, G. and Kingsford, C. 2011, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers, *Bioinformatics*, **27**, 764–70.
23. Hirakawa, H., Shirasawa, K., Kosugi, S., et al. 2014, Dissection of the octoploid strawberry genome by deep sequencing of the genomes of *Fragaria* species, *DNA Res.*, **21**, 169–81.

24. Li, R., Zhu, H., Ruan, J., et al. 2010, *De novo* assembly of human genomes with massively parallel short read sequencing, *Genome Res*., **20**, 265–72.
25. Grabherr, M.G., Haas, B.J., Yassour, M., et al. 2011, Full-length transcriptome assembly from RNA-Seq data without a reference genome, *Nat. Biotechnol*., **29**, 644–52.
26. Li, B. and Dewey, C.N. 2011, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, *BMC Bioinform*., **12**, 323.
27. Kent, W.J. 2002, BLAT–the BLAST-like alignment tool, *Genome Res*., **12**, 656–64.
28. Xue, W., Li, J.T., Zhu, Y.P., et al. 2013, L_RNA_scaffolder: scaffolding genomes with transcripts, *BMC Genom*., **14**, 604.
29. Price, A.L., Jones, N.C. and Pevzner, P.A. 2005, *De novo* identification of repeat families in large genomes, *Bioinformatics*, **21**, i351–8.
30. Trapnell, C., Pachter, L. and Salzberg, S.L. 2009, TopHat: discovering splice junctions with RNA-Seq, *Bioinformatics*, **25**, 1105–11.
31. Trapnell, C., Roberts, A., Goff, L., et al. 2012, Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, *Nat. Protoc*., **7**, 562–78.
32. Li, H., Handsaker, B., Wysoker, A., 1000 Genome Project Data Processing Subgroup., et al. 2009, The sequence alignment/map format and SAMtools, *Bioinformatics*, **25**, 2078–9.
33. Parra, G., Bradnam, K. and Korf, I. 2007, CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes, *Bioinformatics*, **23**, 1061–7.
34. Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. 2015, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, *Bioinformatics*, **31**, 3210–2.
35. Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M. and Stanke, M. 2016, BRAKER1: unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS, *Bioinformatics*, **32**, 767–9.
36. Lomsadze, A., Burns, P.D. and Borodovsky, M. 2014, Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm, *Nucleic Acids Res*., **42**, e119.
37. Stanke, M. and Waack, S. 2003, Gene prediction with a hidden Markov model and a new intron submodel, *Bioinformatics*, **19**, ii215–25.
38. Quevillon, E., Silventoinen, V., Pillai, S., et al. 2005, InterProScan: protein domains identifier, *Nucleic Acids Res*., **33**, W116–20.
39. Llorens, C., Futami, R., Covelli, L., et al. 2011, The Gypsy Database (GyDB) of mobile genetic elements: release 2.0, *Nucleic Acids Res*., **39**, D70–4.
40. Eddy, S.R. 2009, A new generation of homology search tools based on probabilistic inference, *Genome Inform*., **23**, 205–11.
41. Lowe, T.M. and Eddy, S.R. 1997, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, *Nucl. Acids Res*., **25**, 955–64.
42. Huson, D.H. and Scornavacca, C. 2012, Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks, *Syst. Biol*., **61**, 1061–7.
43. Tatusov, R.L., Fedorova, N.D., Jackson, J.D., et al. 2003, The COG database: an updated version includes eukaryotes, *BMC Bioinform*., **4**, 41.
44. Li, L., Stoeckert, C.J.J. and Roos, D.S. 2003, OrthoMCL: identification of ortholog groups for eukaryotic genomes, *Genome Res*., **13**, 2178–89.
45. Kurtz, S., Phillippy, A., Delcher, A.L., et al. 2004, Versatile and open software for comparing large genomes, *Genome Biol*., **5**, R12.
46. Cheng, X., Jiang, H., Zhao, Y., et al. 2010, A genomic analysis of disease-resistance genes encoding nucleotide binding sites in *Sorghum bicolor*, *Genet. Mol. Biol*., **33**, 292–7.
47. Stanke, M., Diekhans, M., Baertsch, R. and Haussler, D. 2008, Using native and syntenically mapped cDNA alignments to improve de novo gene finding, *Bioinformatics,* **24**, 637–44.
48. Kajitani, R., Toshimoto, K., Noguchi, H., et al. 2014, Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads, *Genome Res*., **24**, 1384–95.
49. Tanaka, Y., Sasaki, N. and Ohmiya, A. 2008, Plant pigments for coloration: anthocyanins, betalains and carotenoids, *Plant J*., **54**, 733–49.
50. Mochida, K., Sakurai, T., Seki, H., et al. 2017, Draft genome assembly and annotation of *Glycyrrhiza uralensis*, a medicinal legume, *Plant J*., **89**, 181–94.
51. Sakai, M., Hirata, H., Sayama, H., et al. 2007, Production of 2-phenylethanol in roses as the dominant floral scent compound from L-phenylalanine by two key enzymes, a PLP-dependent decarboxylase and a phenylacetaldehyde reductase, *Biosci. Biotechnol. Biochem*., **71**, 2408–19.
52. Chen, X.M., Kobayashi, H., Sakai, M., et al. 2011, Functional characterization of rose phenylacetaldehyde reductase (PAR), an enzyme involved in the biosynthesis of the scent compound 2-phenylethanol, *J. Plant Physiol*., **168**, 88–95.
53. 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.
54. Scalliet, G., Piola, F., Douady, C.J., et al. 2008, Scent evolution in Chinese roses, *Proc. Natl. Acad. Sci. USA*, **105**, 5927–32.
55. Muhlemann, J.K., Klempien, A. and Dudareva, N. 2014, Floral volatiles: from biosynthesis to function, *Plant. Cell Environ*., **37**, 1936–49.
56. Auldridge, M.E., McCarty, D.R. and Klee, H.J. 2006, Plant carotenoid cleavage oxygenases and their apocarotenoid products, *Curr. Opin. Plant Biol*., **9**, 315–21.
57. Theßen, G. 2001, Development of floral organ identity: stories from MADS house, *Plant Biol*., **4**, 75–85.
58. Mibus, H., Heckl, D. and Serek, M. 2011, Cloning and characterization of three *APETALA1/FRUITFULL* like genes in different flower types of *Rosa* × *hybrida* L, *J. Plant Growth Regul*., **30**, 272–85.
59. Kitahara, K., Hirai, S., Fukui, H. and Matsumoto, S. 2001, Rose MADS-box genes ′MASAKO BP and B3′homologous to class B floral identity genes, *Plant Sci*., **161**, 549–57.
60. Hibino, Y., Kitahara, K., Hirai, S. and Matsumoto, S. 2006, Structural and functional analysis of rose class B MADS-box genes ′MASAKO BP, euB3, and B3′: Paleo-type AP3 homologue ′MASAKO B3′ association with petal development, *Plant Sci*., **170**, 778–85.
61. Kitahara, K. and Matsumoto, S. 2000, Rose MADS-box genes ′MASAKO C1 and D1′ homologous to class C floral identity genes, *Plant Sci*., **151**, 121–34.
62. Gion, K., Suzuri, R., Ishiguro, K., et al. 2012, Genetic engineering of floricultural crops: Modification of flower colour, flowering and shape, *Acta Hortic*., **953**, 209–16.
63. Matsumoto, S. and Kitahara, K. 2005, MADS-box genes in rose: expression analyses of AGAMOUS, PISTILLATA, APETALA3 and SEPALLATA homologue genes in the green rose, *Acta Hortic*., **690**, 203–10.
64. Dubois, A., Raymond, O., Maene, M., et al. 2010, Tinkering with the C-Function: a molecular frame for the selection of double flowers in cultivated roses, *PLoS One*, **5**, e9288.
65. Iwata, H., Gaston, A., Remay, A., et al. 2012, The TFL1 homologue KSN is a regulator of continuous flowering in rose and strawberry, *Plant J*., **69**, 116–25.
66. van Doorn, W.G. and Kamdee, C. 2014, Flower opening and closure: an update, *J. Exp. Bot*., **65**, 5749–57.
67. Ma, N., Xue, J., Li, Y., et al. 2008, Rh-PIP2; 1, a rose aquaporin gene, is involved in ethylene-regulated petal expansion, *Plant Physiol*., **148**, 894–907.
68. Yamada, K., Takahashi, R., Fujitani, C., et al. 2009, Cell wall extensibility and effect of cell-wall-loosening proteins during rose flower opening, *J. Japan. Soc. Hort. Sci*., **78**, 242–51.
69. Xue, J., Yang, F. and Gao, J. 2009, Isolation of Rh-TIP1; 1, an aquaporin gene and its expression in rose flowers in response to ethylene and water deficit, *Postharvest Biol. Technol*., **51**, 407–13.
70. Dai, F., Zhang, C., Jiang, X., et al. 2012, RhNAC2 and RhEXPA4 are involved in the regulation of dehydration tolerance during the expansion of rose petals, *Plant Physiol*., **160**, 2064–82.
71. Chen, W., Yin, X., Wang, L., et al. 2013, Involvement of rose aquaporin RhPIP1; 1 in ethylene-regulated petal expansion through interaction with RhPIP2; 1, *Plant Mol. Biol*., **83**, 219–33.

72. Xiang, Y., Huang, C. H., Hu, Y., et al. 2017, Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication, *Mol. Biol. Evol.*, **34**, 262–81.
73. Yu, C., Luo, L., Pan, H., Guo, X., Wan, H. and Zhang, Q. 2014, Filling gaps with construction of a genetic linkage map in tetraploid roses, *Front. Plant Sci.*, **5**, 796.
74. Hibrand-Saint Oyant, L., Crespel, L., Rajapakse, S., Zhang, L. and Foucher, F. 2008, Genetic linkage maps of rose constructed with new microsatellite markers and locating QTL controlling flowering traits, *Tree Genet. Genomes*, **4**, 11–23.
75. Zhang, L.H., Byrne, D.H., Ballard, R.E. and Rajapakse, S. 2006, Microsatellite marker development in rose and its application in tetraploid mapping, *J. Am. Soc. Hort. Sci.*, **131**, 380–7.