nature genetics

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41588-018-0110-3

In the format provided by the authors and unedited.

The *Rosa* genome provides new insights into the domestication of modern roses

Olivier Raymond[1,18], Jérôme Gouzy[2,18,19], Jérémy Just[1,18], Hélène Badouin[2,3,18], Marion Verdenaud[1,4,18], Arnaud Lemainque[5], Philippe Vergne[1], Sandrine Moja[6], Nathalie Choisne[7], Caroline Pont[8], Sébastien Carrère[1], Jean-Claude Caissard[6], Arnaud Couloux[5], Ludovic Cottret[2], Jean-Marc Aury[5], Judit Szécsi[1], David Latrasse[4], Mohammed-Amin Madoui[5], Léa François[1], Xiaopeng Fu[9], Shu-Hua Yang[10], Annick Dubois[1], Florence Piola[11], Antoine Larrieu[1,17], Magali Perez[4], Karine Labadie[5], Lauriane Perrier[1], Benjamin Govetto[12], Yoan Labrousse[12], Priscilla Villand[1], Claudia Bardoux[1], Véronique Boltz[1], Céline Lopez-Roques[13], Pascal Heitzler[14], Teva Vernoux[1], Michiel Vandenbussche[1], Hadi Quesneville[7], Adnane Boualem[4], Abdelhafid Bendahmane[4], Chang Liu[15], Manuel Le Bris[12], Jérôme Salse[8], Sylvie Baudino[6], Moussa Benhamed[4,19], Patrick Wincker[5,16,19] and Mohammed Bendahmane[1,19]*

[1]Laboratoire Reproduction et Développement des Plantes, Univ Lyon, ENS de Lyon, UCB Lyon 1, CNRS, INRA, Lyon, France. [2]LIPM, Université de Toulouse, INRA, CNRS, Castanet-Tolosan, France. [3]Univ Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR5558, Villeurbanne, France. [4]Institute of Plant Sciences Paris-Saclay (IPS2), CNRS, INRA, University Paris-Sud, University of Evry, University Paris-Diderot, Sorbonne Paris-Cite, University of Paris-Saclay, Orsay, France. [5]CEA–Institut de Biologie François Jacob, Genoscope, Evry, France. [6]Univ Lyon, UJM-Saint-Etienne, CNRS, Saint-Etienne, France. [7]UR1164–Research Unit in Genomics-Info, INRA, Université Paris-Saclay, Versailles, France. [8]INRA/UBP UMR 1095 Genetics, Diversity and Ecophysiology of Cereals, Clermont-Ferrand, France. [9]Key Laboratory of Horticultural Plant Biology, College of Horticulture & Forestry Sciences, Huazhong Agricultural University, Wuhan, China. [10]Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing, China. [11]Univ Lyon, Université Claude Bernard Lyon 1, CNRS, ENTPE, UMR5023 LEHNA, Villeurbanne, France. [12]Aix Marseille Université, Avignon Université, CNRS, IRD, IMBE, Institut Méditerranéen de Biodiversité et d'Ecologie, Marseille, France. [13]INRA, US 1426, GeT-PlaGe, Genotoul, Castanet-Tolosan, France. [14]Institut de Biologie Moléculaire des Plantes, CNRS, UPR 2357, Strasbourg, France. [15]Center for Molecular Biology (ZMBP), University of Tübingen, Tübingen, Germany. [16]CNRS, Université d'Evry, UMR 8030, Evry, France. 17Present address: Centre for Plant Sciences, Faculty of Biological Sciences, University of Leeds, Leeds, UK. 18These authors contributed equally: Olivier Raymond, Jérôme Gouzy, Jérémy Just, Hélène Badouin, Marion Verdenaud. 19These authors jointly supervised this work: Mohammed Bendahmane, Jérôme Gouzy, Moussa Benhamed, Patrick Wincker. *e-mail: mohammed.bendahmane@ens-lyon.fr

Supplementary Information

The Rosa genome provides new insights into the domestication of modern roses

Raymond *et al*.

Supplementary Information Guide

- **Supplementary Information.pdf**: This document includes :

- 14 sections of Supplementary Notes that contain methods, results, illustrations and discussion (Rose history; Homozygous rose line production; Genome sequencing and assembly; Genome annotation, Rose epigenome; Paleogenomics and ancestry of the rose genome, Structure of diversity in *Rosa* species including resequencing of divers rose cultivars and species; analyses of gene pathways involved in scent and color, as well as genes involved in auxin pathway, flowering genes and gamete formation; Supplementary references).

- 23 Supplementary Figures

- 9 Supplementary Tables

- **Supplementary Data 1.xlsx:** This table contains the correspondence of Homozygous and Heterozygous (sets of alleles) annotation.

- **Supplementary Data 2.xlsx:** This table contains the principal component analyses (PCA) data used for the structuration of diversity in rose genotypes along the seven chromosomes

- **Supplementary Data 3. xlsx:** This table reports the results of biochemical analyses of Volatile Organic Compounds (VOCs)

- **Supplementary Data 4.xlsx**: This table contains the manual annotation of genes related to scent and gene expression.

- **Supplementary Data 5.xlsx:** This table reports the results Predicted Auxin Response Factor gene family of *R. chinensis*.

- **Supplementary Data 6:** This file contains the multiple alignments of protein sequences from *Rosa chinensis* and *Arabidopsis thaliana* for 8 MADS-box gene families (fasta format).

- **Supplementary Data 7.xlsx:** This table contains the correspondence of homozygous and heterozygous scent genes annotation and identification of the most likely allele copies.

- **Supplementary Data 8.xlsx:** This table contains the correspondence of homozygous and heterozygous MYB genes annotation and identification of the most likely allele copies with *in silico* expression patterns in different rose tissues.

- **Supplementary Data 9.fasta:** This file contains the predicted TPS gene family of *R. chinensis* and other species.

- **Supplementary Data 10.fasta:** This file contains the predicted MYB10 and related MYBs in *R. chinensis, Malus and Fragaria.*

The Rosa genome provides new insights in the design of modern roses

Raymond *et al*.

Supplementary Notes

Content

1. Rose history

The genus *Rosa* represents a group of plants that appears to have undergone extensive reticulate evolution with interspecific hybridization, introgression and polyploidization. These evolutionary processes have led to the emergence of traits that respond to humankind's hedonistic expectations and have represented an incredible source of diversity. Rose domestication is a particularly complex model produced by hundreds of years of breeding and is based on altering whole pathways and networks. Rose domestication happened at least twice independently in ancient China and the peri-Mediterranean area, encompassing part of Europe and the Middle East[1,2]. In these two regions, generations of rose breeders had fastidiously selected the most desirable traits of *Rosa* species by meticulous observation. Ornamental features, therapeutic and cosmetic values have certainly motivated the domestication in these two world areas. Crosses between *Rosa* species and cultivars have created complex polyploid cultivars that exhibited the most advantageous parent's traits such as recurrent flowering, good looking flowers, pleasant scent, cold hardiness and pathogens resistance[1,3,4]. Two biological groups are particularly important: the Damask roses cultivated for the production of oils and fragrance and the Chinese roses that were unique in their continuous flowering.

The Chinese rose *R. chinensis* is among the few species that participated in breeding programs. In China, roses have been cultivated for a very long time, dating back to the reign of Chin-Nun (2737-2697 BC)[2]. The earliest cultivated Chinese roses were bred from local indigenous forms that grew wild in the mountains of China, probably in the Yunnan and Sichuan areas[5,6]. The second steps in the evolutionary history of the rose is the encounter of the two genic pools from the 18th century that led to the introgression of the continuous flowering, a trait from the Chinese roses in the occidental rose genome. Since the 19th century, massive controlled hybridization allowed the creation of numerous varieties. *R. chinensis* is considered among the most important species that participated in the subsequent extensive hybridization using the gene pools from the European / Mediterranean / Middle-East (mostly tetraploid) and Chinese (mostly diploid) roses. These processes engendered the parental cultivars of the modern-day roses (modern rose cultivars, *Rosa* x *hybrida*)[1,7]. These hybridizations likely happened independently several times and produced triploid hybrids. Supposedly, the production of unreduced gametes allowed breeders to retrieve fully fertile tetraploid hybrids and overcome this triploid block. One of the major Chinese roses used in the creation of modern roses was 'Old Blush' (Parson's Pink China), which also transmitted the recurrent flowering character. Yet, *R. chinensis* 'Old Blush' displays specific phenotypical traits that pinpoint a possible hybrid origin. We generated a high-quality genome sequence of *R. chinensis* 'Old Blush' and we resequenced rose species and/or cultivars that could help in understanding the hybrid architecture of 'Old Blush'. Moreover, our resequencing effort aimed to capture an image of the diversity that is at the origin of the modern-day *R.* x *hybrida* complex genotypes, as well as the allopolyploid origins of *R. gallica* and *R. damascena*. Since the species involved in domestication and later hybridization / introgression events mostly belong to Synstylae, Chinenses and Cinnamomeae sections, our resequencing effort was focused on them to reflect their diversity. Finally, in order to describe the genomic reorganization resulting from the combination of tetraploid European and diploid Asian genomes after hybridization or introgression, we resequenced the emblematic *R.* x *hybrida* 'La France'. Bred in 1867 in Lyon, France by the Guillot family, *R.* x *hybrida* 'La France' is the first modern rose Tea hybrid cultivar[8] that combines growth vigor traits from European species and recurrent blooming from Chinese species. Supplementary Table 2 (below) describes the list of the genotypes that were resequenced in this work along with their ploidy levels and site of sampling.

2. Production of homozygous rose line derived from heterozygous *Rosa chinensis* 'Old Blush'

2.1 Methods

R. chinensis 'Old Blush' plants were grown in a greenhouse at 25°C/19°C day/night temperature with 16 h/8 h day/night supplemental light provided by sodium vapor and metal halide bulbs. Flower buds (Supplementary Fig. 1a) were sampled when the majority of microspores were at the mid-late uninucleate/early bicellular developmental stages (Supplementary Fig. 1b-e) and then surface-sterilized with Pursept® A Xpress for 1 minute, followed by a treatment with a bleach solution (1.5 % active chlorine) containing 0.5% Tween 20 for 15 minutes. Buds were then rinsed 4 times with sterile de-ionized water. Anthers were aseptically dissected from buds, and microspores were isolated as described[9] with the first centrifugation being performed at 100 g for 3 minutes, followed by two centrifugations at 65 g for 3 minutes. Microspores were then suspended in B medium[10], pH 6.5. Microspore viability was checked by FCR test[11] and the developmental stage was assessed by DAPI staining[12] (Supplementary Fig. 1b-e). In all experiments, the microspores viability was around 50%. Density was then adjusted to 100,000 microspores/mL and the suspension was pretreated at 4°C in darkness for 21 days in Falcon 353001 Petri dishes sealed with Parafilm® (1.5 mL microspore suspension per dish). Microspores were then rinsed twice with cold B medium and centrifuged at 50 g for 3 min at 4°C. A portion of 160,000 microspores from this fraction was then suspended in 600 µL of AT12 medium corresponding to AT3 medium[9] supplemented with 4.5 µM 2,4-D and 0.44 µM BAP, pH 5.8, and then incubated in a 12-well plate sealed with Parafilm® at 25°C in the dark. After 3 weeks, the medium was carefully replaced with 600 µL of fresh AT12 medium, and the culture was further incubated with the same conditions. Developing micro-calli (*ca.* 0.5 mm diameter) were observed about 8 weeks after subculture (Supplementary Fig. 1f). Developing micro-calli were isolated and subcultured individually in 300 µL of the same medium in a 24-well plate sealed with Parafilm® in the same conditions. After 2 weeks, calli were plated onto a CM3 solid medium containing MS salts[13], B5 vitamins[14], 30 g/L sucrose, 2.5 mM MES, pH 5.8, supplemented with 4.5 µM 2,4D, 0.44 µM BAP and 6.5 g/L VitroAgar (Kalys Biotechnologie, Saint Ismier, France). After 7 weeks of culture, developing calli were subcultured once on CM3 medium for 12 weeks. At this stage, several calli issued from the same experiment of microspore culture displayed developing somatic embryos (Supplementary Fig. 1g). Embryogenic calli were propagated by repeated subcultures, every 4-6 weeks, on callus maintenance medium[15] or by repeated subcultures on embryo maintenance medium EMM[16]. Homozygosity of developing embryogenic calli was assessed using High Resolution Melting (HRM) analyses and by observing the *k*-mer spectrum of Illumina reads derived from this homozygous material. HRM analyses were performed with the Applied MeltDoctor TM HRM master mix (ThermoFisher Scientific), following the manufacturer's instructions, using the following primer pairs known to amplify heterozygous loci in 'Old Blush' genome: RC008174_F TGCAACTGGCTTTGAGGTTG, RC008174_R AACCACTGGGCCAAACAAAG, RC008432_F ACGCAGCTGAAATGTATGGC, RC008432_R TCTTCTTGCAGCTCCGTTTC, RhEF1-QS1 GGGTAAGGAGAAGGTTCACATC, RhEF1-QAS1 CAGCCTCCTTCTCAAACCTCT. To regenerate homozygous rose plantlets, embryo cotyledons taken from calli propagated on EMM were processed as described[16].

2.2 Results

Roses exhibit high heterozygosity levels that hamper high quality genome assembly. To overcome this difficulty, we developed a protocol that allows 'Old Blush' microspores to switch from gametophyte to sporophyte development. We used a combination of fine-tuning a starvation medium, cold stress and hormonal treatments to induce microspores that initiate divisions and to form cell clusters (Supplementary Fig. 1f) after about 11 weeks of culture. Clusters were developed and yielded both embryogenic and proliferating calli that were then maintained on various media (Supplementary Fig. 1g,h).

DNA genotyping (HRM) of isolated calli showed that all tested loci were homozygous (Supplementary Fig. 1k). Developing calli displayed the same homozygous profile indicating that they likely derived from a unique microspore development event. This callus was designated *R. chinensis* HzRDP12 (hereafter *Rc*HzRDP12; Supplementary Fig. 1g,h). The *k*-mer spectrum of Illumina reads derived from *Rc*HzRDP12 provided the final proof that the genome of *Rc*HzRDP12 genome was homozygous, demonstrating a loss of heterozygosity in 'Old Blush' (Supplementary Fig. 1l). Experiments exploring the potential of *Rc*HzRDP12 material have revealed that it is possible to maintain the embryogenic capacity of produced calli through several subcultures. Furthermore, we readily regenerated plantlets with normal morphological phenotype from *Rc*HzRDP12 somatic embryos (Supplementary Fig. 1i).

To determine the ploidy level of the homozygous *Rc*HzRDP12 material, we performed fluorescence-activated cell sorting (FACS) analysis. We used *R. chinensis* 'Old blush' leaves, cultivated *in vitro*, as control. Nuclei were isolated from *Rc*HzRDP12 calli or from young leaves of regenerated plantlets, as previously described[17], and stained by adding 1 µg/mL DAPI (Sigma) for 1 hour at room temperature. FACS analyses were performed using MACSQuant VYB (Miltenyi Biotec) cytometer and analyzed by FlowJo software (FlowJo LLC). One major peak corresponding to diploid (2N) cells was observed after DAPI staining for *Rc*HzRDP12 (Supplementary Fig. 1j). The ploidy profile of this homozygous material was identical to that of the heterozygous *R. chinensis* 'Old Blush' plants, used as a control. In all samples, the majority of cells were diploid and low proportion of polyploid cells (4N and 8N), frequently observed in young tissues, was detected. These data demonstrate that haploid cells originating from the homozygous callus did undergo spontaneous genome duplication during regeneration resulting in diploid homozygous *R. chinensis* 'Old blush' callus and plant material.

To the best of our knowledge, this is the first demonstration of the production of a homozygous rose plantlet. The use of such approach opens possibilities to implement haplomethods in rose genetics and breeding. This possibility to generate Recombinant Inbred Like materials paves the way for novel breeding strategies in roses, *e.g.* F1 breeding or reverse breeding. With respect to more fundamental research, availability of homozygous rose genotypes may foster the study of a number of processes in simpler genetic models (*e.g.* developmental mechanisms or metabolic pathways). In particular, homozygous genotypes represent promising models for functional genetics.

3. Genome assembly

3.1 Genome sequence assembly

3.1.1 Meta-assembly process

The first generation of long-read genome assembly software such as PBcR[18] and FALCON[19] enabled the assembly of chromosomes or chromosome arms of small or medium sized genomes[18,20]. The genome assemblies of genomes with higher repeat complexity (*e.g.* plant genomes) were still composed of several hundreds or thousands of contigs[20,21] and required code modifications to adapt overlap filtering to peculiarities of complex genomes[20]. Recently, CANU has revisited the detection of spurious edges in the graph of overlaps by introducing filtering parametrization at the read level leading to more accurate and contiguous assemblies[22]. Nonetheless, two CANU assemblies of 80x PacBio data of the *R. chinensis* genome generated around 400 contigs and the other metrics varied depending on the number of corrected reads used (Supplementary Fig. 2a). To circumvent this difficulty and improve assembly contiguity, we developed a companion software called til-r for editing the FALCON overlap graphs by defining local cut-offs for each read end (Supplementary Fig. 2c and next section). We ran FALCON/til-r with stringent cut-offs to generate four alternate assemblies (Supplementary Fig. 2a) expecting that additional and "difficult" gaps would be resolved. Then, we used CANU to perform a meta-assembly of our six primary assemblies in which the number of contigs ranged between 298 and 413 and an N50 between 3.37 and 7.95 Mb. As the CANU version 1.4 was unable to handle such large sequences, primary assemblies were transformed into very long overlapping sequences with a maximum of 100 kb (50 kb overlap) prior the meta-assembly. The meta-assembly was executed with a minimal overlap of 10 kb and the overlap based trimming step was activated in order to trim spurious contigs ends (found in one assembly only). The meta-assembly is composed by only 82 contigs for an N50 of 24 Mb (Supplementary Fig. 2a) showing the complementarity of primary assemblies. The obtained assembly with a few contigs was then easily integrated with a high-density map as already described in the main text and in the Online Method section.

3.1.2 The til-r software

til-r is a C software implementing heuristics that aim to filter the graph of overlaps generated by the FALCON pipeline. It replaces the call to the program "fc_ovlp_filter" in the script "run_falcon_asm.sub.sh" in FALCON version 0.7.

The different pipeline functions, inputs and outputs, and defaults parameters are described in Supplementary Fig. 2c. The four heuristics, the assumptions or combinations of assumptions on which they are based and how they are applied at the read-end level are presented here:

Assumption #1: an overlap that spans a non-repeated region is not ambiguous. The length of PacBio reads is long enough to span a large majority of repeated regions.

Heuristic #1: a list of non-repeated regions can be provided to til-r as a tabular text file or automatically computed. Only overlaps spanning a non-repeated region are considered. To quickly identify likely non-repeated regions in reads, we first randomly sub-sample the read dataset to obtain less than 1x coverage per slice. All reads are classified in one slice. The number of slices is computed depending on genome size and

targeted coverage. In each slice, the corresponding overlap positions are used to define repeated regions. After consolidating of repeated regions over all slices, the list of non-repeated regions is defined.

Assumption #2: The identity percentage for overlaps depends on the read end quality and some tolerance must be allowed for trying to avoid dead ends (read ends without any overlaps above the cut-off). At a given identity cut-off, the overlaps list contains true positive overlaps but also false positives in the case of repeated regions in the genomic regions. The identity percentage for the false positive overlaps is expected to be lower than the one for the true positives. The best identity percentage found is an indirect measure of read end quality.

Heuristic #2: A deltapci parameter that permits tuning the maximum difference allowed between the overlap with the best identity percentage overlap and the other overlaps that are taken into account. When the difference is too high, the overlaps are removed even if their identity percentages are above the general cut-off.

Assumption #3: The best overlap graph algorithm selects the largest overlaps to build the path of reads. Read ends that are not accurately corrected can lead to dead ends. For overlaps that span a likely non-repeated region (see Heuristic #1), taking into account the size of the overhang, can help select neighbor reads that permit a minimum span of the genomic region.

Heuristic #3: Reads with dead ends are iteratively removed from the graph until no edit. Remove overlaps where wing size defined as Minimum (overlap length, overhang length) is below a given number of nucleotide cut-off.

Assumption #4: The check of transitive consistency of overlaps can be used to clean up the graph of dubious overlaps.

Heuristic #4: Removing overlaps with reads that do not overlap the best scoring overlap. Removing overlaps kept by only one read of the pair (the reciprocal was removed by previous filters).

The software (source code and amd64 Linux binaries) can be downloaded from http://lipm-bioinfo.toulouse.inra.fr/download/til-r/.

3.2 *Rosa chinensis* homozygous genome Illumina sequencing

We produced 147x of Illumina paired-end and mate pair reads (Supplementary Table 4), following the protocol described in Supplementary Note 4.1. The data were then used for subsequent statistical analyses.

3.3 Pseudo-chromosomes validation using Three-dimensional proximity information (Hi-C)

Methods

About 0.5 g of formaldehyde-fixed leaf tissues were used to prepare 2 independent *in situ* Hi-C libraries. The sample fixation was performed as for ChIP-seq in this study. Nuclei extraction, nuclei permeabilization, chromatin digestion, and proximity ligation treatments were performed essentially as previously described[23]. The extracted nuclei were resuspended in 150 µL 0.5% SDS, split equally into three tubes and incubated at 62°C for 5 min. After which 145 µL water and 25 µL 10% Triton X-100 were added, and incubated at 37°C for 15 min. Next, the nuclei in each tube were digested by adding 25 µL 10x NEB buffer 3 (100 mM NaCl, 50 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM DTT, pH 7.9) and 50 U of DpnII restriction enzyme, and incubated at 37°C overnight. The next day, the nuclei were incubated at 62°C for 20 min to inactivate the restriction enzyme. Next, the digested chromatin was blunt-ended by adding 1 µL of 10 mM dTTP, 1 µL of 10 mM dATP, 1 µL of 10 mM dGTP, 25 µL of 0.4 mM biotin-14-dCTP, 14 µL water and 4 µL (40 U) Klenow fragment, and incubated at 37ºC for 2 hr. Subsequently, 663 µL water, 120 µL 10x blunt-end ligation buffer (300 mM Tris-HCl, 100 mM $MgCl_2$, 100 mM DTT, 1 mM ATP, pH 7.8), 100 µL 10% Triton X-100, and 20 Weiss U T4 DNA ligase were added to start proximity ligation. The ligation reaction was placed at room temperature for 4 hr. After ligation, the nuclei were collected by centrifugation at 1,000 rcf for 3 min, and then resuspended in 750 µL SDS buffer (50 mM Tris-HCl, 1% SDS, 10 mM EDTA, pH 8.0), and incubated with 200 µg proteinase K at 55°C for 30 min. The formaldehyde crosslink was reversed by adding 30 µl 5M NaCl to the solution followed by overnight incubation at 65°C. The recovery of Hi-C DNA and subsequent DNA manipulations were performed as described previously[24]. The final libraries were sequenced on an Illumina NextSeq instrument with 2 x 75 bp reads.

Results

Over the past few years, three-dimensional proximity information obtained by Hi-C was reported as an efficient method to construct spatial proximity maps of many eukaryotes to help assemble their genomes[25]. We constructed spatial proximity maps of the rose genome using chromosome conformation capture sequencing (Hi-C) at a resolution of 400 kb and then used it to evaluate and confirm the genome assembly and the rose 7 pseudo-chromosomes constructions. The two Hi-C-libraries (denoted A and B, with respectively 198,638,690 and 219,337,784 reads) were independently analyzed with Hi-C-Pro pipeline (default parameters and LIGATION_SITE=GATCGATC)[26]. Reads were first cut for adaptors with trim_galore software[27] and then independently aligned against the genome (bowtie2, end-to-end algorithm[28]) in a 2-steps protocol to avoid chimeric reads. Only valid ligation products were kept independently for the two libraries (26,067,262 and 23,907,222 respectively, for lib A and lib B) then merged together for the

interaction map construction. The genome was divided into equally sized bins and number of contacts observed between each pair of bins, was reported. Finally contact maps were plotted with HICPlotter software [29]. The high collinearity between the genetic map based pseudomolecules anchoring (Figure 1) and Hi-C based contact map information corroborated the overall assembly quality.

3.4 Localization of centromeres

Centromeric repeats are expected to have a very conserved length, with sequence variations. To localize biological centromeres, first we detected tandem repeats (TRs) genome-wide using the TRF software[30], with parameters "2 7 7 80 10 80 2000 –d –m –l 16", and obtained 11,069 TR motifs. We used Blastn with parameters "M=2 N=-5 Q=7 R=7 E=1e-10 wordmask=none filter=none V=10000000 B=10000000" to count the number of occurrences of each TR pattern on the genome (Supplementary Fig. 10a). We selected patterns of an over-represented length in the genome (lengths: 61-65, 75-80, 92-97, 115-118, 159-162, 175-176, 522-526, 1044-1053), that were then assembled into contigs by length, with Cap3 [31]. We obtained 931 contigs that we mapped on the genome using Blastn, with parameters "M=1 N=-1 Q=2 R=2 E=1e-10 V=2147483647 B=2147483647 gapS2=500 gapX=500 kap". 108 contigs that had more than 1,000 occurrences in the genome, were kept. We then visually inspected the distribution of their sequence coverage along the pseudomolecules by looking for TR highly repeated localized in a narrow region of each chromosome, with a strong anti-correlation with gene density, and a correlation with TE density. We selected 13 TR motifs of 61-65, 92-97 and 159-162 in length. Their combined density along the genome (shown in Supplementary Fig. 10b), allowed to localize the centromere for each chromosome.

4. Sequencing and assembly of heterozygous *R. chinensis* 'Old Blush' genome

4.1 Library preparation and sequencing

Four Illumina PE libraries (overlapping and tightly sized PE libraries) were prepared using a semi-automated protocol. Two independent DNA fragmentations were performed from the extracted DNA using the E210 Covaris instrument (Covaris, Inc., USA) to generate fragments mostly around 300 bp (for the overlapping library) or 600 bp (for the library with 3 insert sizes of 500 bp, 600 bp, and 800 bp) (Supplementary Table 5). End repair, A-tailing and Illumina compatible adaptors (BioScientific, Austin, TX, USA) ligation were performed using the SPRIWorks Library Preparation System and SPRI TE instrument (Beckmann Coulter), according to the manufacturer protocol.

DNA fragments were then PCR-amplified using Platinum Pfx DNA polymerase (Invitrogen) and Illumina adapter-specific primers. Fragments of around 300 bp were size selected on 3% agarose gel while fragments of around 500 bp, 600 bp and 800 bp were selected on 2% agarose gel. Library traces were validated on an Agilent 2100 Bioanalyzer (Agilent Technologies, USA) and quantified by qPCR using the KAPA Library Quantification Kit (Kapa Biosystems) on a MxPro instrument (Agilent Technologies, USA). The PE libraries were sequenced using 100 base-length read v3 chemistry in paired-end flow cell on the Illumina HiSeq 2000 (Illumina, USA).

The Mate Pair libraries were prepared using the Nextera Mate Pair Sample Preparation Kit (Illumina, San Diego, CA). Briefly, genomic DNA (4 µg) was simultaneously enzymatically fragmented and tagged with a biotinylated adaptor. Tagged fragments were size-selected (3-5; 5-8 and 8-11 Kb) through regular gel electrophoresis, and circularized overnight with a ligase. Linear, non-circularized fragments were digested and circularized DNA was fragmented to 300-1000 bp size range using Covaris E210. Biotinylated DNA was immobilized on streptavidin beads, end-repaired, then 3'-adenylated, and Illumina adapters were added. DNA fragments were PCR-amplified using Illumina adapter-specific primers and then purified. Finally, libraries were quantified by qPCR and library profiles were evaluated using an Agilent 2100 bioanalyzer (Agilent Technologies, USA). Each library was sequenced using 100 base-length read chemistry on a paired-end flow cell on the Illumina HiSeq 2000 (Illumina, USA) (Supplementary Table 5).

4.2 Heterozygous *R. chinensis* 'Old Blush' genome assembly

4.2.1 Assembly

We used ALLPATHS-LG (version 44837) on all the read libraries listed in Supplementary Table 6, except the 8-11 kb MP library. At the contiguing stage, we obtained 104,181 assembly graphs (contigs with ambiguities), spanning 746.5 Mb (Supplementary Table 6). Around 0.55% of the total contig length is represented as ambiguities, and more than 93.8% of these ambiguities have exactly two forms. We believe these ambiguities represent the residual polymorphism between haplotypes, for the fraction of the genome that hasn't been resolved in two distinct haplotypes. After scaffolding, we obtained an assembly of 882.7 Mb (Supplementary Table 6).

4.2.2 Validation of assembly completeness and separation of haploptypes

The assembly sequence was assessed with BUSCO v3.0.2b[32] which found 1,346 complete gene models out of 1440 (93.5%) and 28 fragmented (1.9%); 73.8% of complete genes are in more than one copy, while this is the case for only 4.5% of the homozygous genome. We mapped the 80,714 rose transcripts from[33] with Blastn (parameters: "E=1e-8 W=9 wordmask=dust links hspsepSmax=12000") and est2genome[34]. Supplementary Fig. 11 displays the distribution of the number of matches depending of the applied identity percent cutoff. We found that at 90% sequence identity cut-off, 76.9% of transcripts have at least one match, and around 71.5% among them have exactly two matches. Along with the overall heterozygous assembly length (882.7 Mb, for an estimated haploid size of 560 Mb), these results show that our assembly process managed to discriminate the two alleles for around 70% of the genes.

4.3 Localization of crossing-overs on *Rc*HzRDP12 genome

The homozygous *R. chinensis* *Rc*HzRDP12 genotype was obtained from microspores culture (Extended Notes 2) and therefore underwent a meiosis. To identify putative loci of crossing-overs that occurred during meiosis, we mapped Illumina reads from 5 distinct libraries from the heterozygous genome (paired-ends 370 bp, 480 bp and 630 bp, mate-pairs 3.3 kb and 5.4 kb; Supplementary Table 6) on the constructed pseudo-chromosomes and we counted pairs in which only one read had a match, in 10 kb-long windows. Normalization was made using the number of consistent pairs for each library. We observed 50 windows with over-represented one-end mapped pairs in at least two libraries. They were then kept as candidate crossing-over loci (indicated as horizontal dashed lines on Supplementary Fig. 12, yellow frame).

To validate this strategy, we looked for breakpoints in the sequence conservation with genotypes related to the inferred parents and close genotypes of 'Old Blush' (See below Supplementary Notes 8). We cut single reads of length 100 bp in the reads obtained from *R. wichurana*, *R. gigantea*, *R. chinensis* 'Spontanea', *R. chinensis* 'Old Blush', *R. odorata* 'Hume's Blush', *R. chinensis* 'Sanguinea', *R.* x *hybrida* 'La France' (see below Supplementary Notes 8) and mapped them on *R. chinensis* *Rc*HzRDP12 genome sequence with Smalt (http://www.sanger.ac.uk/science/tools/smalt-0, v0.7.6), with sequence similarity cutoffs of 99%, 98% and 97%. We counted mapped reads over 200 kb windows, and normalized in each window with the number of homozygous *R. chinensis* reads mapped in the same conditions, to estimate sequence conservation between the 8 genotypes and the homozygous *R. chinensis*. The outcome is shown on Supplementary Fig. 12, with

red lines of three different intensities depicting the three similarity cutoffs. Conservation can be higher than 1 at a low stringency due to repeated sequences.

The observed segmental conservation pattern was in accordance with the inferred close relationship of the genotypes. Moreover, the opposite patterns of conservation with WIC and GIG/SPO (high conservation with one genotype and low conservation with the other genotype) confirmed that the haplotype extracted in *Rc*HzRDP12 is a mosaic of genomes closely related to the sequenced WIC and GIG/SPO, thus confirming the hybrid origin of 'Old Blush'. Six candidate crossing-overs perfectly co-localized with breakpoints in the conservation between the homozygous and heterozygous *R. chinensis* genomes or with inferred parents. It is to note that crossing-overs that happened in regions where the two haplotypes of the heterozygous genome have the same relative conservation with WIC and GIG/SPO could not be confirmed by this method.

Conservation between homozygous and heterozygous *R. chinensis* genomes also showed a segmental pattern (Supplementary Fig. 12, OB column), demonstrating that the heterozygosity level of 'Old Blush' is not homogeneous. Moreover, most of the genome length had a conservation value of 0.60-0.75, indicating that, since one of the haplotype of 'Old Blush' was completely identical to the extracted one, only one third of the reads from the other haplotype could match *Rc*HzRDP12 genome sequence. This estimate of one third of matching reads was consistent with the lowest values observed in WIC and GIG/SPO. One region of chromosome 3 (29.2-49.2 Mb) had a conservation value of 1, indicating that both haplotypes were completely identical to the homozygous one. Two smaller regions (chr2:34.0-47.2 Mb and chr4:63.2-67.0 Mb) were also nearly homozygous, at a lesser extent. Conservation with the inferred relatives of 'Old Blush' (HUM, SAN MUT and FRA) showed a more fragmented pattern, suggesting that they underwent more crosses. The homozygous region on chromosome 3 of 'Old Blush' is shared with HUM, SAN and FRA, but not with WIC nor GIG/SPO, suggesting that this region could have been selected during modern rose breeding.

4.4 Cytoplasmic origin of *Rosa chinensis* 'Old Blush'

To get more insight into the origin of 'Old Blush', we used the mapping of reads from *R. chinensis* 'Old Blush', *R. wichurana*, *R. gigantea*, *R. chinensis* 'Spontanea' on the homozygous *R. chinensis Rc*HzRDP12 genome (Supplemental Notes 4.3) to infer the most probable cytoplasmic origin of *Rosa chinensis* 'Old Blush'. After applying a cutoff at 100% identity (whole read length) on the read alignments, we computed the length of chloroplast genome covered by reads. Reads from 'Old Blush' were covering 98.941% of chloroplast genome (mean depth of coverage=11,286), reads from SPO were covering 98.323% of it (DC=3,924), while reads from WIC and GIG were covering only 95.706 and 95.037% of it, respectively (DC=3,351 and 2,247), meaning that among the inferred parents of 'Old Blush', the most probable cytoplasmic origin is *R. chinensis* 'Spontanea'.

5. Genome annotation

5.1 Transcriptome data used for the prediction of gene models

Transcriptome data were generated from *R. chinensis* cultivars floral buds[35] grown in a greenhouse with the following conditions: 16 h / 8 h day/night and 25°C / 14°C day/night temperature, as described previously. RNA preparation was performed as previously described[35]. RNA integrity was checked using Nano chip, Agilent 2100 Bioanalyzer (Agilent Technologies, Waldbroon, Germany) and then used to generate 3' cDNA library for Illumina sequencing (GATC Biotech) according to the manufacturers protocols (Illumina). Adapters were clipped using cutadapt[36] and regions with an average Phred quality lower than 28 in average along a 4 bp sliding window were trimmed using custom scripts based on BioPerl[37]. Reads shorter than 25 bp after trimming and unpaired reads (in the case of paired-end sequencing) were discarded. Read counts after trimming ranged from 19 to 325 millions. The above RNAseq data were combined with RNA-seq data from other organs of *R. chinensis* 'Old Blush' described in[33] and RNA-seq data from *R. chinensis* 'Pallida' a cultivar closely related to 'Old Blush' and from *R. chinensis* 'Viridiflora'[38].

5.2 Annotation of protein coding genes

Gene models were predicted using a fully automated pipeline egn-ep (http://eugene.toulouse.inra.fr/Downloads/egnep-Linux-x86_64.1.4.tar.gz) that manages probabilistic sequence model training, genome masking, transcript and protein alignments computation, alternative splice sites detection and integrative gene modelling by the EuGene software (release 4.2a[39]). Four protein databases were aligned (blastx[40]) to contribute to translated regions detection: i) TAIR10 [41] ii) Swiss-Prot - December 2015 iii) a plant subset of Uniprot proteins – December 2015 and iv) the proteome of *Brachypodium distachyon* release 192[42]. Proteins similar to REPBASE[43] were removed from datasets prior to alignment. Chained alignments spanning less than 50% of the length of the database protein were removed. The Illumina-based RNAseq datasets described in 5.1 were assembled with an iterative *k*-mer strategy based on velvet[44], parameters: -cov_cutoff 4 -read_trkg yes -exp_cov 100 -min_contig_lgth 150 -max_divergence 0.05 -long_mult_cutoff 0) allowing a homogenous integration of RNAseq data with two additional public datasets of Sanger, 454 and unigene sequences (Genbank January 2015[45]. The four sets of "expressed sequence tags" were aligned on the genome using gmap[46] and only the best scoring hit was kept. Spliced alignments spanning at least 30% of the EST sequence length at a minimum of 97% identity were retained. In case of splicing ambiguity, the introns with the highest number of occurrences in the four datasets were selected. Repeat masked loci (Red -len 16[47]) were unmasked by hits with EST databases, TAIR or *B. distachyon*. The gene modeling algorithm used the standard EuGene 4.2a parameters, except that non-canonical GC/donor sites were allowed and transcribed regions longer than 200nt without any predicted CDS were reported as ncRNA. Other ncRNA genes were predicted by tRNAScan-SE (tRNAs[48], RNAMMER (RDNAs[49]) and rfamscan (Rfam release 12[50]. After removing redundant ncRNA predictions, 45,469 protein-coding genes and 4,918 non-protein-coding genes were annotated. The set of predicted peptides was run on the BUSCO plant/embryophyta_odb9 release 2[32] and 1,389 complete plus 23 fragmented gene models out of a total of 1,440 (96.5% and 1.5% respectively) were detected. This automatic annotation was post-processed to remove gene models overlapping the annotation of transposable element leading to a final set of 36,377 protein-coding gene models.

EuGene pipeline was used to annotate the heterozygous genome of 'Old Blush' with the same sources of evidences, leading to a set of 61,908 protein-coding gene models. The set of predicted mRNAs was assessed with BUSCO plant/embryophyta_odb9 v3.0.2b[32] which found 1,351 complete gene models out of 1,440

(93.8%) and 47 fragmented (3.3%). 73.4% of complete genes were in more than one copy (5.0% in the homozygous genome), indicating we recovered the two alleles of a majority of the genes.

To determine allele pairs, we compared with Blastp (parameters: "W=3 Q=7 R=2 matrix=BLOSUM90 B=500 V=500 E=1e-15 hitdist=60 hspsepqmax=10 hspsepsmax=10 sump") the complete proteomes from *Rosa chinensis* homozygous and heterozygous, *Fragaria vesca* v1.0 and v2.0.a1[51], *Rubus occidentalis*[52], *Malus* x *domestica* v1.0[53], and GDDH13 v1.1[54], *Pyrus communis*[55], *Pyrus bretschneideri*[56], *Prunus mume*[57], *Prunus persica*[58], *Ziziphus jujube* cv. 'Dongzao'[59] and cv. 'Junzao'[60], *Medicago truncatula*[61,62], *Junglans regia*[63], *Populus trichocarpa*[64], *Carica papaya*[65], *Arabidopsis thaliana* TAIR10[41], *Vitis vinifera* V1[66], *Lycopersicon esculente* v2.3[67], *Oryza sativa* cv. 'Japonica' v1.0.31[68] and *Brachypodium distachyon* v.3.1[42]. For each *Rosa chinensis* predicted protein, we only kept its matches with *Rosa* proteins bidirectional and better than any match against another species. We then looked for cliques in the graph of alignments, defining them as putative "allele sets". Most of the allele sets contains one gene model from the homozygous genome, and two from the heterozygous genome (10,148 out of 27,287; Supplementary Table 7), corresponding to the canonical case where the two alleles have been resolved in the heterozygous genome. 7,813 alleles sets contain one homozygous and one heterozygous gene models, corresponding to cases where alleles were assembled as a consensus in the heterozygous genome. Other cases could be due to gene duplications and/or gene losses having occurred independently in the two haplotypes, or to residual transposable elements in our gene annotation.

By aligning the complete nucleotide sequence of genes predicted in one assembly on the genome sequence of the other assembly with Blastn (parameters: "M=1 N=-3 Q=3 R=3 E=1e-30 wordmask=dust hspsepSmax=30 hspsepQmax=30 links sump") and looking for overlaps between matches and genome annotation, we built a correspondence table between genes from the two genomes, provided as Supplementary Data 1.

5.3. tRNA and rRNA annotation

Transfer RNA genes were predicted using tRNAScan-SE v1.3[48] with parameters "-t R -C". Only predictions with scoring higher than 20 were kept. We obtained 757 predicted tRNA genes, and 114 predicted pseudogenes. 1,153 tRNA genes and 155 pseudogenes were predicted in the heterozygous genomes. Ribosomal RNA genes were predicted using RNAmmer v1.2 (RDNAs[49]), with eukaryotic parameters set for nuclear chromosomes and bacterial parameter for organellar chromosomes. We obtained 313 predicted rRNA genes. Most of them were on chromosomes 1 (149 genes) or 3 (123 genes). 49 rRNA genes were predicted in the heterozygous genomes.

5.4. Transposable elements and repeats annotation

5.4.1 De novo transposable element annotation

We used the REPET package (https://urgi.versailles.inra.fr/Tools/REPET) to produce a genome-wide annotation of repetitive sequences on the homozygote PacBio genome (7 pseudo-chromosomes and 46 unassigned contigs) and the heterozygote Illumina genome (15,938 scaffolds) (see Online Methods). In this genome, the most abundant TE fraction is retrotransposons also called class I elements (31.6%) and in particular, Long Terminal Repeat retrotransposons (LTR-RTs) represent 22.9% with Ty3/Gypsy superfamily being more abundant than Ty1/ Copia superfamily. Non-LTR retrotransposons (LINE and potential SINE)

contribute approximately to 7% and class II elements (DNA transposons and Helitrons) to 11.6%. The 22% remaining correspond respectively to unclassified repeats (7.87%), chimeric consensus with two classifications (7.51%) and to potential host genes repeated in this genome (around 6%). These genes were identified and kept in this study. We also identified 2,765 caulimoviridae insertions, representing 1.25 of the genome (Supplementary Fig. 4a,b; Supplementary Table 1).

Finally, we used this library of 3,933 consensuses to annotate the TEs copies in the heterozygote Illumina genome assembly (15,938 scaffolds). Each consensus has at least one copy on the heterozygote genome and the global and non-redundant TE content in the final annotation was 54.7% based on 746 Mb of sequence assembly excluding undefined bases (Ns). The TE families distribution in this genome is the same as in the homozygote with some difference for the Ty3/Gypsy superfamily (9.8%), class I-LARD elements (0,7%) and chimeric (4.4%) (Supplementary Fig. 4a,b; Supplementary Table 1).

6. The first Rose epigenome

6.1 ChIP-seq assay

ChIP assays were performed using anti-H3K9ac (Millipore, ref. 07-352) or anti-H3K27me3 (Millipore, ref. 07-449) antibodies, using a procedure adapted from[69]. Briefly, petals at the onset of flower opening were fixed in 1% (v/v) formaldehyde. Petal tissues were homogenized and nuclei were isolated and lyzed. Cross-linked chromatin was sonicated using a water bath Bioruptor UCD-200 (Diagenode, Liège, Belgium) (30s/30s on/off pulses, at high intensity for 60 min). Protein/DNA complexes were immunoprecipitated with antibodies, overnight at 4°C with gentle shaking, and incubated for 1h at 4°C with 50 µL of Dynabeads Protein A (Invitrogen, Ref. 100-02D). The beads were washed 2 × 5 min in ChIP Wash Buffer 1 (0.1% SDS, 1% Triton X-100, 20mMTris-HCl pH 8, 2 mM EDTA pH 8, 150 mMNaCl), 2 × 5 min in ChIP Wash Buffer 2 (0.1% SDS, 1% Triton X-100, 20 mMTris-HCl pH 8, 2 mM EDTA pH 8, 500 mMNaCl), 2 × 5 min in ChIP Wash Buffer 3 (0.25 M LiCl, 1% NP-40, 1% sodium deoxycholate, 10 mMTris-HCl pH 8,1 mM EDTA pH 8) and twice in TE (10 mMTris-HCl pH 8, 1 mM EDTA pH 8). ChIPed DNA was eluted with two 15 min incubations each at 65°C with 250 µL Elution Buffer (1% SDS, 0.1 M $NaHCO_3$). Chromatin was reverse-crosslinked by adding 20 µL of 5 M NaCl and incubated over-night at 65°C. Reverse-cross-linked DNA was submitted to RNase and proteinase K digestion, and extracted with phenol-chloroform. DNA was ethanol precipitated in the presence of 20 µg of glycogen and resuspended in 20 µL of nuclease-free water (Ambion) in a low-bind DNA tube. Ten nanograms of IP or input DNA was used for ChIP-Seq library construction using NEB-Next Ultra II DNA Library Prep Kit for Illumina (New England Biolabs) according to manufacturer's recommendations. Ten PCR cycles were used for all libraries. The library quality was assessed with Agilent 2100 Bioanalyzer (Agilent), and the libraries were subjected to high-throughput sequencing by NextSeq 500 (Illumina).

6.2 ChIP-Seq bioinformatics analysis

Preprocessing of sequenced reads for quality was performed using FASTQC[70]. A single end library H3K27me3 and a paired end library H3K9ac and theirs corresponding inputs were cleaned and trimmed with trim_galore[27] with following parameters: mean Phred quality score greater than 20 ; read length greater than 10 after trimming ; retain unpaired reads. Remaining reads were aligned onto the *R. chinensis* genome with bowtie2[28] with a maximum mismatch of 1 bp and unique mapping. Result files were formatted with samtools[71] and coverage calculated with Picard tools[72]. To determine the target regions of H3K9ac ChIP-Seq, the Model-based Analysis of ChIP-Seq (MACS2)[73] was used (number of duplicate reads at a location:1; nandwidth:300; mfold of 5:30; q-value cutoff:0.05). SICER was used to detect H3K27me3 modification regions SICER was used (window size:200, gap size:600)[74]. HOMER[75] was used to associate H3K9ac peaks were located into a -2kb;+1kb windows around the gene TSS. To associate H3K27me3 genes, bedtools intersect[76] was used to keep genes that are overlapped with a H3K27me3 region. Genes and mark densities were calculated using Rstudio [RStudio Team] and plotted with Rstudio and Circos[77] for circular visualization. The average coverage profile along the genic region and 1 kb gene flanking region was plotted using NGSplot[78] To cluster the H3K9ac and H3K27me3 peaks, linear normalization and clustering of tag density with Density Array method (window size 50 bp; 2 kb gene flanking region) was performed using SeqMINER[79].

6.3 Results

Genome-wide studies in plants have provided evidence for the role of H3K9ac and H3K27me3 in gene activation and repression, respectively[80-84]. The roles of these histone modifications in rose remain unknown and represent a represent a limitation to the full understanding of how thousands of bioprocesses are regulated. To determine the genomic landscape of these marks, we performed a ChIP-seq analysis using H3K9ac and H3K27me3 antibodies on petals from a heterozygous plant. A minimum of 17 millions of mapped reads was obtained (Supplementary Fig. 13a). The MACS2 and SICER algorithms, which are designed to detect sharp and broader histone peaks, respectively[73,74], were used to determine loci that are significantly enriched with H3K9ac or with H3K27me3 (Supplementary Fig. 13a,b). We identified 23,770 H3K9ac marked genes and 11,223 H3K27me3 marked genes for homozygous genome; 28726 H3K9ac marked genes and 15850 H3K27me3 marked genes for heterozygous genome (Supplementary Fig. 13b).

Next, we analyzed the distributions of the two histone marks at the chromosome and gene levels. To analyze the genome wide distribution, we used the homozygous assembly. However, in order to capture both haplotypes diversities, all gene level analysis were performed on heterozygous assembly. At the chromosomal scale, we observed an enrichment of both marks in gene-rich regions, which is consistent with the role of these histone marks in the control of gene expression (Supplementary Fig. 13c,d). In order to detail the H3K9ac and H3K27me3 distributions at the gene level, the peaks obtained for both modifications were analyzed. We found that the peak length of H3K9ac ranged from 400 bp to 800 bp (Supplementary Fig. 13e), located preferentially at the TSS regions, (Supplementary Fig. 13f). In contrast, H3K27me3 peaks presented an averaged length that ranged from 4,000 bp to 8,000 pb, covering the entire gene body (Supplementary Fig. 13e,g). Those patterns were consistent with previous studies on different plant species, highlighting conserved aspects of the epigenetic system in the plant kingdom[85]. As expected, integration of H3K9ac and H3K27me3 data sets showed an anti-correlation between those two marks (Supplementary Fig. 13h). Altogether, these results show that in rose, as in other plant species, H3K9ac and H3K27me3 are distributed along the gene body, supporting the role of these two marks in gene regulation.

To connect H3K9ac and H3K27me3 histone marks with gene expression, we generated and integrated RNA-seq data. We confirm that H3K9ac and H3K27me3 in rose are associated with gene expression and gene repression, respectively (Supplementary Fig. 13k). Genes that are associated with both marks show an intermediate expression profiles. To determine if the level of acetylation or methylation could be correlated with gene expression, we equally divided all the genes into four groups, based on their expression levels. Then we plotted them on their H3K9ac or H3K27me3 profile (Supplementary Fig. 13i,j). We observed that H3K9ac level increases with expression level while H3K27me3 showed the opposite pattern, where it displayed a high enrichment in the lowest-expressed genes. These results suggest that in rose the more a gene is marked by H3K9ac and H3K27me3, the more it will be expressed and repressed, respectively.

7. Rosaceae genome evolution for translational research

In order to assess the paleohistory of *R. chinensis* within the *Rosaceae* family, we performed a comparative genomic investigation of *Rosa* with apricot (*Prunus mume*[57]), peach (*Prunus persica*[58]), apple (*Malus domestica*[53]), pear (*Pyrus bretschneideri*[56]) and strawberry (*Fragaria vesca*[51]), using the genome alignment parameters and ancestral genome reconstruction methods described in Salse 2016[86]. Conserved gene adjacencies deliver an ancestral *Rosaceae* karyotype (ARK) consisting of 9 protochromosomes (or Conserved Ancestral Regions, CARs) with 8861 protogenes (Supplementary Fig. 5a, top). The complete dot-plot based deconvolution into nine reconstructed CARs of the observed synteny and paralogy between ARK and the investigated species validate the nine proposed protochromosomes as the origin of *Rosaceae* (Supplementary Fig. 5a, bottom). Our evolutionary scenario, reconciling the modern genome structures to the founder ARK, clearly established that apricot and peach emerged from an ancestral *Prunoideae* karyotype (APK) structured in 8 protochromosomes (with 16333 protogenes) deriving from ARK through 2 ancestral chromosome fissions and 4 fusions. The duplication of ARK followed by at least 11 ancestral chromosome fissions and 12 fusions, shaped the ancestral *Maloideae* karyotype (AMK) in 17 protochromosomes (with 12,634 protogenes), as the founder ancestor of the modern apple and pear genomes[53], while no similar duplication was found in *Rosa* or *Fragaria* genomes. Finally, the ancestral *Rosoideae* karyotype (ARoK) of the modern strawberry and rose genomes, structured into 8 protochromosomes with 13,070 protogenes, derived from ARK through one ancestral chromosome fission and 2 fusions. While the modern strawberry genome experienced an extra ancestral chromosome fusion from ARoK to reach its modern genome structure, rose genome went through one fission and 2 fusions, independent from strawberry, to reach its modern genome structure. Our comparative genomics-based evolutionary scenario unravels the *Rosaceae* paleohistory from the reconstructed ancestral *Rosaceae* karyotype (ARK) with 9 protochromosomes and 8,861 protogenes delivering the complete catalog of paralogous and orthologous gene relationships between the modern *Rosaceae* genomes as well as the reconstructed ancestor (ARK, APK, AMK, ARoK). The gained knowledge can now be used as a guide to perform translational research between the six-investigated species to accelerate the dissection of conserved agronomical traits (Supplementary Fig. 5a, bottom).

Rosoideae radiative evolution: The relative phylogenic relationships between rose, raspberry and strawberry, all from the Rosoideae subfamily, are currently weakly supported, due to a lack of molecular data[87,88]. The hypothesis is that *Rosa* and *Fragaria* diverged more recently from one another than from *Rubus*. We used our rose genome sequence, and that of *Rubus occidentalis*[52] and *Fragaria vesca*[51] to address this question, using *Malus* x *domestica*[54] as an outgroup.

We selected the 748 genes that were identified as complete and in unique copy in the four genomes with BUSCO plant/embryophyta_odb9 dataset[32] (v3.0.2b). Based on their coding sequences, we computed 748 individual trees, using MUSCLE v3.8.31[89] and PhyML v3.1[89] with parameters "0 I 1 1000 HKY e e 4 e BIONJ y y". We observed that 61.5% of the trees had a bootstrap value of 996/1000 or more. Among them, 68.7% support the hypothesis of a shorter distance between *Rosa* and *Fragaria*, compared with *Rubus* (Supplementary Fig. 5b, barplot). The consensus tree obtained from the concatenation of 600 gene CDSs with the same method, with an additional step using Gblocks v0.91b[90] showed the same tendency (Supplementary Fig. 5b, bottom right). However, by plotting the *Rosa-Fragaria* and *Rosa-Rubus* phylogenetic distances gene by gene (Supplementary Fig. 5b, dot plot in lower panel), we observed that the

dots followed the diagonal (in blue) and that the slope was only marginally different from 1 (5% confidence interval in red). These results favor the hypothesis that the three genera diverged approximately at the same time, suggesting a process of evolutionary radiation inside the Rosoideae subfamily.

Despite being evolutionary close to each other, *Rosa* and *Fragaria* have differing genome size, respectively 560 and 240 Mb. We retrieved 3 datasets of genomic reads from distinct *Fragaria vesca* subspecies from NCBI (SRR1513870, SRR1513871 and 1513872) to compare the fraction of repeated *k*-mers to the one of our *Rosa* sequencing data. Individual reads were cleaned, and regions with a Phred quality lower than 26 in average along a 4 bp sliding window were trimmed. Reads shorter than 55 bp were discarded. We filtered out reads matching *R. chinensis* 'Old Blush' chloroplastic or mitochondrial genomes, or *Fragaria vesca* genome (NC_015206[51]), using Bowtie v1.1.1[69]. We randomly subsampled *Rosa* datasets to 2.4 Gb to have a similar size to *Fragaria* ones (repeated 10-16 times). We used Jellyfish v2.2.6[91] to count *k*-mers of length 55, 47 and 43 bp. We considered a *k*-mer as over-represented when it was seen more than 5 times its expected occurrence count, estimated for the genome size and the depth of coverage of the dataset. We observed that 6.4 to 7.8% of the genome of *Fragaria vesca* is represented by repeated *k*-mers (Supplementary Fig. 5c), while this fraction ranges from 8.6 to 15.6% for *Rosa* spp., with a mean around 11%. This result suggested that most of the genome size difference could be explained by the relative richness in repeats.

8. Structure of diversity in *Rosa* species

8.1 Methods

During rose breeding, cultivars have been obtained by inter-specific crosses and backcrosses, then maintained by vegetative multiplication. Thus, a limited number of meiosis and recombination events occurred. We assumed that the size of the introgressed fragments should be large in the genomes or sub-genomes of hybrid rose cultivars, in contrast with what could be observed if hybridization events were followed by extensive sexual reproduction.

The reference genome is a double haploid obtained from a single meiosis event of the hybrid cultivar *R. chinensis* 'Old Blush'. If the density of variants for a given resequenced genotype in a genomic interval is a function of the distance between the haplotype of *R. chinensis* 'Old Blush' in the reference genome and each haplotype or subgenome of the resequenced genotype, discrete levels of variant density along the genome could indicate either genomic regions that have different introgression histories or different haplotypes of the heterozygote *R. chinensis* 'Old Blush' in the double haploid reference genome (limits would correspond to crossing-overs, with an expected number of one per chromosome).

Discrete variations of variant density can therefore be used to segment the genome into regions that may have different introgression histories. As we were interested in the history of hybridization between the Chinenses section on the one hand, and the Synstylae or Cinnamomae sections, we took into account the resequenced genotypes of hybrid cultivars related to *R. chinensis* (*R. chinensis* 'Mutabilis', *R. chinensis* 'Sanguinea', *R. odorata* 'Hume's Blush'), as well as the triploid hybrid cultivar, *R.* x *hybrida* 'La France', also related to the Chinenses section. Variant density was computed by sliding windows of 1 Mb for each resequenced genotype. We cut the genome at positions corresponding to inflexion points in the density of variants in at least one hybrid cultivar. This resulted in a segmentation in 35 genomic segments, ranging from 2 to 56 Mb.

DNA purification and sequencing: Leaf material was collected from 14 *Rosa* species and cultivars grown at the ENS-Lyon-France, at the Lyon Botanical Garden, France, at "Jardin Expérimental" at Colmar, France or at a private rose garden (O. Masquelier, La Bonne Maison, Lyon, France) (Supplementary Table 2). Approximately 100 mg of young leaves were ground in liquid nitrogen using mortar and pestle. No previous nuclei purification step was undertaken, but ground samples were collected in 1.5 mL of homogenization buffer (Tris 15 mM, EDTA 2 mM, NaCl 20 mM, KCl 80 mM, pH 8.5) with 0.7% (W/V) PVP40, 0.5% (V/V) Triton X100 and 0.1% (V/V) 2-mercaptoethanol. Samples were homogenized for 1h by centrifugation at 20 cycles / min and pellets were retrieved by 20 min centrifugation at 3000 g. Genomic DNA was then extracted using DNeasy Plant kit (Qiagen, MD, USA). DNA integrity was inspected via gel electrophoresis (0.7% agarose) and total DNA was quantified by fluorometry using Picogreen® (Applied Biosystems/Life Technologies, Carlsbad CA, USA).

DNAseq libraries were constructed and sequenced at Génoscope-Evry-France or at Eurofins Genomics, Ebersberg, Germany. Paired-end sequenced DNA libraries were constructed using Illumina's TruSeq DNA LT kit following the manufacturer's recommendations. The genomic DNA libraries were sequenced on the Illumina HiSeq2500 (2 x 100) platform using the HiSeq SBS Kit v4 sequencing chemistry (Illumina).

8.2 Results

8.2.1 Reads mapping, SNP calling and filtering

Illumina paired-end reads of the four *Rosa* species with read lengths greater than 100 nt were mapped to the reference genome with the GLINT software (http://lipm-bioinfo.toulouse.inra.fr/download/glint/), with the following parameters: --no-lc-filtering --best-score --mate-maxdist 10000 --lmin 80 --mmis 16 --step 2. The mismatch cut-off was increased to 24 for the ten *Rosa* species with read lengths equaling 150 nt.

Variants were called for each genotype with SAMtools mpileup[71] and Varscan[92], with the following parameters for low coverage genotypes: min-coverage=5, min-reads2=5, --min-avg-qual 15, min-var-freq=0.1 --p-value 0.01 and with more stringent parameters for the high coverage 'Old Blush' heterozygous genotype: --min-coverage 50 --min-reads2 25 --min-avg-qual 15 --min-var-freq 0.1 --p-value 0.01. Variants with a mapping coverage higher than 60 and 300 in the fourteen resequenced *Rosa* species and in the *R. chinensis* 'Old Blush' genotype respectively, were filtered out.

8.2.2 Origin of the *Rosa chinensis* 'Old Blush' genotype

The section Chinenses comprises old cultivated Chinese roses that are supposed to result from crosses between two wild species, *R. gigantea*, and *R. chinensis* 'Spontanea', a rare wild species[2]. One of the first Chinese roses used in the creation of modern roses, transmitting the recurrent flowering character was 'Old Blush' (= Parson's Pink China). *R. gigantea* and *R. chinensis* 'Spontanea' have single flowers, entire stipules and free short styles[93], but this first cultivated recurrent flowering Chinese rose exhibits branched flower heads, free but protruding styles and dentate stipules. These morphological traits could indicate a close relationship to section Synstylae roses. Section Synstylae is characterized by branched flower heads, pectinate or dentate stipules and styles connate in a slender column, exerting a flat and conical discus. Phylogenetic studies based on molecular data have shown that the Synstylae are allied to the Chinenses[94]. To identify the parents of 'Old Blush' and the origin of the fragments in the double homozygote 'Old Blush' reference genome, we computed the density of homozygote and heterozygote variants in 1 Mb sliding windows in the resequenced genomes of *R. chinensis* 'Spontanea' and *R. gigantea* for the Chinenses section, *R. moschata*, *R. wichurana* and *R. arvensis* for the Synstylae section, as well as the heterozygote *R. chinensis* 'Old Blush' genotype (Supplementary Fig. 14). Discrete variants density levels could be observed. In *R. gigantea*, very low values (< 1 homozygote variant per kb) were observed in around 28% of the genome, corresponding to regions of the double homozygote 'Old Blush' originating from *R. gigantea* or a very closely related species (Supplementary Fig. 14c). Such low values were not observed in the resequenced genotypes of the Synstylae section genotypes (Supplementary Fig. 14e-g), nor in *R. chinensis* 'Spontanea', which thus appears as a lesser contributing ancestor of 'Old Blush' ancestor (Supplementary Fig. 14d). This is corroborated by the data in Supplementary Note 4.4 indicating that although a genotype closer to the sequenced *R. chinensis* 'Spontanea has transmitted its cytoplasm to 'Old Blush', the latter's genome is closer to *R. gigantea* than to *R. chinensis* 'Spontanea'. Furthermore, a region extending from 30 to 46.5 Mb on chromosome 3 and originating from the Chinenses section displayed a very low variant density (< 2 variants per kb), but displayed normal mapping coverage (Supplementary Fig. 12), and is therefore homozygous in the *R. chinensis* 'Old Blush' heterozygote genotype (Supplementary Fig. 14b). Our analysis confirms that *R. gigantea* or a close relative is a parent of *R. chinensis* 'Old Blush'. Furthermore, principal component analyses indicate that diversity is structured along certain chromosomic regions according to patterns that are

intermediate between those of true Synstylae and Chinenses species (Supplementary Fig. 6, fragments 7.1 ; 7.2 ; 6.1 ; 4.4 ; 4.3 ; 4.2 ; 1.4). This is consistent with the hypothesis of a hybrid origin of 'Old Blush' and raises the question about the identity of its second progenitor in the Synstylae section.

Principal component analyses also highlight the origin of the tetraploid *R. gallica* and *R. damascena*. These two cultivars appear intermediate between the Synstylae and the Cinnamomeae sections (Supplementary Fig. 6), although closer to the Synstylae section, which suggests a hybrid Synstylae x Cinnamomeae origin.

9. Rose scent gene pathways

Modern roses have inherited scent from both European and Chinese lineages through many manmade crosses. The diverse fragrances are linked to the expression of the different enzymatic pathways inherited from wild species. Rose scent compounds belong to 3 major classes, terpenoids, benzenoids/phenylpropanoids and fatty acid derivatives. In contrast to the extensive literature on the chemistry of rose scent, very few studies have dealt with scent production in rose petal cells[95]. Identifying enzymes responsible for the biosynthesis of major scent compounds and their transcription regulatory pathways have thus become major goals in rose research.

9.1. Methods

9.1.1 Biochemical analyses of scent composition in roses

We performed a biochemical analysis of scent compounds in the rose genotypes *R. chinensis* 'Old blush', *R. gigantea*, *R. damascena*, *R. gallica*, *R. moschata* and *R. wichurana* that exhibit different scent compositions spanning the rose scent compound diversity. To extract volatile organic compounds (VOCs), petals or stamens were weighed and mixed with hexane in a 1:2 ratio, for 48 h at 4 ºC. Camphor was used as internal standard to estimate compound quantities. Hexane fraction for each sample, was separated, filtered, concentrated and stored at -20°C until analysis in a gas chromatograph coupled to a mass spectrometer (Agilent 6850). Two µL of each sample were injected at split mode with a 2:1 ratio. The injector and detector temperatures were at 250°C and 280 ºC, respectively. The global run time was recorded in ei-mode (35-450 *m/z* mass range) at a scanning rate of 2.94 scan s^{-1}. An electron ionization mass spectrometry (EI-MS) detector operated under an ion source temperature of 23°C and a trap emission current of 35 µA and a 70 eV ionization energy were used. The compounds were separated through a 0.25 mm x 30 m DB-5MS capillary column (J&W Agilent), at a film thickness 0.25 µm, with helium as the carrier gas at a flow rate of 1 mL min^{-1}. The GC oven temperature was programmed to increase from 40°C to 180ºC at rate of 1.50ºC min^{-1}, and from 180ºC to 290ºC at rate of 10ºC min^{-1} and was finally maintained at 290°C for 1 min. All experiments were performed at least two times.

The chromatographic data were analyzed using the Data Analysis software (Agilent) and the volatile substances were identified by screening the WILEY 275, NIST 08, and CNRS libraries for comparison of MS spectra. The Kovats retention indexes (KI) of each substance were calculated using injection data for a homologous set of *n*-alkane (C_8-C_{20}) according to the Kovats formula[96]. Mass spectra similarities combined with KI were then used for compound identification. Concentrations were calculated by comparing of the camphor area to the internal standard[97].

9.1.2 Manual annotation of genes related to scent

The content of VOCs highlights the biochemical pathways that are expressed in *R. chinensis* 'Old Blush' flowers. First, we searched for putative genes acting in each of rose scent pathways by BLAST searches using the heterozygous rose genome. Since only few genes have a known function in rose scent biosynthesis, we used genes sequences from others plant species (*Arabidopsis thaliana*, *Petunia hybrida*, *Fragaria vesca*, *Cucumis melo*…). Secondly, genes corresponding to important biochemical pathways absent in *R. chinensis* 'Old Blush' petals, but present in other rose cultivars, were searched for in *R. chinensis* cv. 'Old Blush'. The rationale is that these later genes may be present but not expressed in *R. chinensis* 'Old Blush'. The method consisted in a BLASTN search in *Rosa* or *Fragaria* entry sequences and a BLASTP search with sequences from other species with the objective to find homologous *R. chinensis* 'Old Blush' genes. The FPKM values

were computed with Tophat[98] and Cufflinks[99] using the rose RNAseq dataset (this work and previously published data[33]).

Perl scripts were used to obtain files corresponding to scaffolds with interesting gene sequences (fasta format), RNA-seq contigs[33] and RNAseq data, and automatic annotation. These files were visualized in Artemis, a genome browser and annotation tool[100]. The transcriptome datasets were used for curation to verify the automatic annotation of each gene sequence, using Artemis. When necessary, a manual annotation was performed to correct the automatic annotation and a new file (gff format) was created. To check the automatic predicted function of each gene sequence studied, a BLASTN search in NCBI using predicted mRNA as the query, was performed and the results were reported in a new file (Genbank format).

The results of this manual annotation with predicted functions are presented in Supplementary Data 4. Genes are organized according to the biosynthetic pathways. For each gene, the FPKM using the EST data[33] and the predicted function by manual annotation and by automatic annotation are given. Generally, more than one sequence corresponded to one Blast query. These sequences were considered as homologous copies of the studied gene, and could be allelic variants or different gene copies. Supplementary Data 7 provides the correspondence between heterozygous IDs and the reference genome annotation (homozygous) and helps identify putative alleles for scent genes.

9.2. Results

The emblematic rose perfume is a bouquet of more than one hundred VOCs, composed of terpenoids, benzenoids/phenylpropanoids, fatty acid derivatives and others chemical families such as fatty acid derivatives or phenolic methyl ethers (PME). The presence and abundance of individual compounds present a wide diversity between species and cultivars. To gain insights into the rose scent composition and diversity in rose, we performed biochemical analyses of major VOCs present in petals of six rose species, *R. chinensis* 'Old blush', *R. gigantea*, *R. damascena*, *R. gallica*, *R. moschata* and *R. wichurana* (Supplementary Data 3). We identified 61 major compounds belonging to the main enzymatic pathways known in roses. Modern roses have inherited scent from both European and Chinese lineages through many manmade crosses. The diverse fragrances are linked to the expression of the different enzymatic pathways inherited from wild species. For example, terpenoids and phenylpropanoids can be found in many wild species, but PMEs are only found in species in the Chinenses section (*R. chinensis* and *R. gigantea*).
The enzymatic pathways of the VOCs are only partially known in roses[101-105] and many biochemical steps remain to be discovered. Data mining of the rose genome reveals candidate genes for this perspective. We took advantage of the rose genome to identify and reconstruct the biosynthesis pathways associated with the relevant scent compounds.

9.2.1 Phenolic methyl ethers

Phenolic methyl ethers (PMEs) are found in roses in the Chinenses botanical section, *R. chinensis* and *R. gigantea* and in many of their hybrids in the "tea" and "hybrid tea" groups. Analyses of petal VOCs (Supplementary Data 3) show that *R. gigantea* can synthesize 6.67 µg/g FW of 3,5-dimethoxytoluene (DMT), which produces the "tea odor" and *R. chinensis* 'Old Blush' can synthesize 19.66 µg/g FW of 1,3,5-trimethoxybenzene (TMB). DMT is synthesized by two specific enzymes, orcinol-O-methyl transferases 1 and 2 (OOMT1 and 2), that catalyze the methylations of orcinol, a substrate[106] (Supplementary Fig. 15). TMB is synthetized by three successive methylations of phloroglucinol, the first step being catalyzed by a phloroglucinol-O-methyl transferase (POMT)[107] (Supplementary Fig. 15). The next steps are probably

catalyzed by OOMT1 and OOMT2. The origins of orcinol and phloroglucinol are not well documented. A phloroglucinol synthase has been characterized in brown algae[108] and an orcinol synthase homologous to a bacterial gene has recently been discovered in *Rhododendron dauricum*[109]. These two genes belong to the polyketide synthase (PKS) family.

Homologous genes known to act in the PMEs pathway could be found in the genome of *R. chinensis* 'Old Blush' genome (Supplementary Data 4; Supplementary Fig. 15). One sequence corresponding to *OOMT1* (RcHt_406.5) and to *OOMT2* (*RcHt_S13.10*) are highly expressed in open flower (FPKM>4500). Other sequences that are close to *OOMTs* (*RcHt_S406.17*, *RcHt_S2315*.2) exhibited weak expression levels in flowers (FPKM<40). A gene encoding for POMT (RcHt_S111.5, RcHt_1962.11) is highly expressed in buds and in stamens (FPKM>550). Since *R. chinensis* 'Old Blush' only emits TMB in a trace amount, it is possible that phloroglucinol is methylated in buds and stamens by POMT and is then methylated in open flower by OOMT1 and OOMT2 to synthesize TMB.

Genes homologous to phloroglucinol synthases (*PKS*) were also found in *R. chinensis* 'Old Blush' genome. Five candidate sequences show high expression levels in flowers. Among these genes, three are highly expressed in buds (*RcHt_S332*.2, FPKM = 240) and in stamens (*RcHt_S332.2*, *RcHt_S55.41, RcHt_S412.26*, FPKM = 68 to 1021). These expression patterns correspond to those of *POMT* and thus could represent candidates to study the initial steps in the TMB pathway. The identified PKSs belong to the type III clade, which is involved in the biosynthesis of specialized metabolites corresponding to aromatic polyketides[109]. Two other sequences (*RcHt_S950.24, RcHt_S117.11*), also in the type III clade, are expressed in open flowers, one of which show high expression level (FPKM of 228). This expression pattern does not correspond to the *POMT* pattern, but to that of *OOMT1* and *OOMT2*, and therefore could also represent good candidates for the PME pathway.

9.2.2 Terpenoids

R. chinensis 'Old Blush' petals produce mostly acyclic monoterpenes like geraniol (89.73 µg/g FW) and geranial (28.34 µg/g FW), while other monoterpenes, such as ß-myrcene, geranyl acetate, nerol, neral, and (+/-)-ß-citronellol, are found in much smaller quantities. Interestingly, rose species belonging to different sections, produce different monoterpenes that contribute to their different scent signatures (Supplementary Data 3, Supplementary Fig. 7). For example, *R. damascena* produces high amount of (+/-)-ß-citronellol (102.68 µg/g FW), while *R. damascena* and *R. gallica* produce high levels of nerol (47.51 and 41.84 µg/g FW, respectively).

Rose compounds analyses (Supplementary Data 3) show that rose petals are also the site of sesquiterpenes biosynthesis. We found that germacrene D and δ-cadinene are produced in *R. chinensis* 'Old Blush petals, elemol is produced in *R. gallica* petals, (*E*)-ß-farnesene and (*E,E*)-farnesol are produced in *R. wichurana* petals, and norterpenes (dihydro-ß-ionol) are produced in *R. chinensis* 'Old Blush' and in *R. gigantea* petals.

In plants, the terpene precursors isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP), are synthesized by two pathways: the 2-C-methyl-D-erythritol 4-phosphate (MEP) pathway and the mevalonate (MVA) pathway (Supplementary Fig. 7). This MEP pathway is compartmentalized in plastids and the MVA pathway is compartmentalized in the cytosol. Furthermore, IPP and DMAPP polymerization does lead to the same volatile terpenoids because different prenyl transferases and terpene synthases are active in the plastids and in the cytosol. IPP and DMAPP polymerization leads to C10 monoterpenes and norterpenes via geranyl diphosphate (GPP) and C40 carotenoid synthesis *via* geranylgeranyl diphosphate (GGPP) synthesis in the

plastids, and to C15 sesquiterpenes via farnesyl diphosphate (FPP) synthesis in the cytosol[110]. Nevertheless, some plants have alternative and unique pathways[111]. In roses, sesquiterpenes seem to be synthesized in the cytosol [102] and norterpenes in the plastids[112], much like other plant species. However, in rose acyclic monoterpenes biosynthesis occurs in the cytosol by a noncanonical enzyme named NUDX1[104]. This unusual subcellular localization raises the question of where GPP biosynthesis is localized. To date, no prenyl transferases and very few terpene synthases have been characterized in roses. A Germacrene D synthase (GDS), has been functionally characterized[102] and a putative linalool synthase, LINS, was identified[113]. Moreover, although many genes in the MVA and MEP pathways are well known in many plants, little information is available in roses.

We used the rose genome sequence to identify homologous genes in the MEP pathway (*DXS, DXR, MCT, CMK, MDS, HDS, HDR* and *IDI*) (Supplementary Data 4; Supplementary Fig. 7). We were able to identify and annotate at least two sequences corresponding to each gene in the rose genome, except for *MCT*. These genes are expressed at low levels in rose petals and all genes, except *IDI*, show relatively weak expression in flowers (Supplementary Data 4). Five sequences of the *DXS* gene were annotated (*RcHt_S1378.5, RcHt_S229.6, RcHt_S254.14, RcHt_S734.26, RcHt_S2705.11*) and none of them showed high expression levels in rose flowers. Four sequences were annotated for *HDR* (*RcHt_S190.25, RcHt_S190.23, RcHt_S3257.2*), two sequences for *DXR* (*RcHt_S387.24, RcHt_S2435.2*), *CMK* (RcHt_S736.6, RcHt_S1563.10), *MDS* (*RcHt_S128.3, RcHt_S280.26*) and *HDS* (*RcHt_S20.74, RcHt_S4142.6*) and only one sequence was annotated for *MCT* (*RcHt_S1965.4*). Among these genes in the MEP pathway, only *HDR* (*RcHt_S190.25*; FPKM ranging from 7 to 16) and *IDI* (*RcHt_S1440.14, RcHt_S7123.2*; FPKM from 43 to 79) showed expression in open flowers and in buds, respectively.

Conversely to the MEP pathway genes, the MVA pathway genes (*AACT*, *HMGS*, *HMGR*, *MVK*, *PMK* and *MVD*) showed higher expression levels in the flower (Supplementary Data 4). Except for *AACT* (*RcHt_S481.35*), at least two sequences for each gene were annotated: *HMGS* (*RcHt_S165.36, RcHt_S180.11*); *HMGR* (*RcHt_S370.28, RcHt_S370.29, RcHt_S2387.6, RcHt_S596.10, RcHt_S1321.13, RcHt_S144.13*), *MK* (*RcHt_S107.22, RcHt_S2220.16); PMK (RcHt_S14556.1, RcHt_S5568.2, RcHt_S5493.2*), *MDD* (*RcHt_S596.10, RcHt_S2633.2, RcHt_S2633.*3) and *IDI* (*RcHt_S1440.14, RcHt_S7123.2*) (Supplementary Fig. 7). Three sequences of *HMGR* (*RcHt_S596.10, RcHt_S1321.13, RcHt_S144.13*) presented an expression with FPKM ranging from 50 to 99 in open flowers. *RcHt_S596.10* is flower specific, according to the EST dataset[33].

We found several prenyl transferase candidate genes in the rose genome (Supplementary Fig. 7), but only three sequences were expressed in open flowers (Supplementary Data 4). The farnesyl diphosphate synthase gene (*RcHt_S4398.3*) encoding a prenyl transferase involved in the synthesis of FPP for the production of sesquiterpenes, is expressed during blooming (100 to 278 FPKM). Concerning GPP biosynthesis for monoterpenes production, the putative heterodimeric geranyl diphosphate synthase large subunit (*RcHt_S620.13*) is expressed in flower buds, in open flowers and in stamens (FPKM from 3 to 39), while the small subunit (*RcHt_S998.24*) shows very low expression levels (FPKM from 7 to 16). It is also possible that *RcHt_S620.13* corresponds to a geranyl geranyl diphosphate synthase involved in the carotenoid biosynthesis pathway. It must be noted that the observed low expression of *GPPS* and the low expression of the MEP pathway genes are inconsistent with the high amount of geraniol in *R. chinensis* 'Old Blush'. Thus, the expression data described above raise the probability that in roses, the MVA pathway could be responsible for all prenyl diphosphates biosynthesis, including GPP. Therefore, if our hypothesis is correct, this will add

another specificity of scent biosynthesis in rose, like what we have previously reported for the NUDX1 hydrolase and geraniol biosynthesis[104].

73 sequences corresponding to terpene synthases have been found in the rose genome (Supplementary Data 9) based on the following criteria: the protein sequence was longer than 390 amino acids (except *RcHt_S12415*.1) and it presented at least some of the characteristic structural motives of TPS (DDxD, NSE/DTE, $RR(x)_8W$)[114]. We performed phylogenetic analyses including TPS from other plants, whose functions have been demonstrated *in vitro* (Supplementary Fig. 16). As expected, rose TPS are distributed in the well-known TPS clades. 44 sequences are grouped in the TPS-a clade, suggesting that they are sesquiterpene synthases. Most of the other sequences are distributed in 2 other TPS groups, TPS-b (15 sequences) and TPS-g (8 sequences), which generally contain monoterpene synthases[114]. Only five rose TPS are expressed in flowers (*RcHt_S4142.3, RcHt_S1216.21, RcHt_S1158.3, RcHt_S12415.1, RcHt_S605.34*). Functional studies and enzymatic assays of these five terpene synthases will help unraveling their putative roles in terpene biosynthesis pathway in rose (Supplementary Fig. 7). *RcHt_S605.34*, which corresponds to the previously characterized GDS[102], is highly expressed in open flowers. In the haploid genome, several putative *LINS* (linalool synthase) or *NES* (nerolidol synthase) sequences are clustered on chromosome 5. These genes are not expressed in rose petals.

Genes corresponding to carotenoid cleavage dioxygenases involved in ionones production (*CCD1*, *RcHt_S2152.4*, *RcHt_637.14* and *CCD4*, *RcHt_S10901.1*) have also been found in the genome. *CCD4*, which shows a very high petal expression in petals at blooming and senescent stages, could be involved in dihydro-ß-ionol biosynthesis in *R. chinensis* 'Old Blush' petals.

9.2.3 Green leaf volatiles

Green leaf volatiles (GLVs), which are alpha-linolenic and linoleic acid derivatives, are generally produced in leaves for defense. With our extraction method, *R. chinensis* 'Old Blush' petal extracts contain the highest amounts of GLVs: (*E*)-2-hexenal, (*Z*)-3-hexen-1-ol, (*E*)-2-hexen-1-ol, hexan-1-ol, (*Z*)-3-hexenyl acetate, (*E*)-2-hexenyl acetate, hexyl acetate and hexanal. The most abundant compounds are (*Z*)-3-hexenyl acetate (32.34 µg/g FW) and (*E*)-2-hexenal (28.26 µg/g FW) (Supplementary Fig. 17). *R. wichurana* and *R. gigantea* also produce hexanal and (*E*)-2-hexenal. *R. damascena* petals present only small amounts of hexan-2-ol and (*E*)-2-hexenal. The first steps of GLVs biosynthesis are unknown in roses, but are well studied in other plant leaves, such as *Arabidopsis thaliana* and *Vitis vinifera*[115]. To get insights into the first steps of rose GVL biosynthesis, we used *A. thaliana* and *V. vinifera* gene sequences to identify their putative homologues in the rose genome (Supplementary Fig. 17). Only genes expressed in flowers have been selected. Homologues of the *13LOX*, *HPL*, *IF*, *ADH* and *AAT*, known to encode for proteins that catalyze the different steps in the GLV pathway were searched for in the rose genome. Two copies of putative gene encoding for linoleate 13S-lipoxygenases (13LOX) have been selected for annotation (*RcHt_S289.22, RcHt_S3147.6)*. Hydroperoxide lyase (HPL) belongs to the cytochrome P450 family. The present annotation identified one HPL gene with certainty (*RcHt_S53.46*) and four cytochrome P450 genes showing high expression in open flowers (FPKM from 104 to 228) were retained as candidates (*RcHt_S63.35, RcHt_S698.32, RcHt_S933.2, RcHt_S3768.2*). The gene encoding for hexenal isomerase (IF) was searched for, but no close homologue could be found. IF protein presents a cupin like domain and one candidate gene (*RcHt_S5960.3*) that harbors this domain was identified in the rose genome. The aldehyde isomers are converted into alcohols by alcohol dehydrogenases (ADH). There are many *ADH* candidate genes in the *R. chinensis* 'Old Blush' genome. For

example, one *ADH* gene, which was cloned in *R. rugosa* (KF724973.1), corresponds to *RcHt_S1703.9*. The last step of this pathway corresponds to the acetylation of alcohol compounds by alcohol acyl-transferases (AAT) (Supplementary Fig. 17). One AAT gene was functionally characterized[116]. It corresponds to *RcHt_S420.25* and *RcHt_S2552.2* sequence.

9.2.4 Benzenoids and phenylpropanoids

R. chinensis 'Old Blush' produces only trace amounts of benzenoids and phenylpropanoids in petals. A small amount of 2-phenylethanol is found in stamens (Supplementary Data 3). Nevertheless, 2-phenylethanol (1029.2 µg/g FW) and ß-phenylethyl acetate are found in *R. damascena*, and 2-phenylethanol alone in *R. gallica, R. moschata* and *R. wichurana*. Eugenol and methyl-eugenol are found in *R. gigantea* and *R. damascena*, while *R. moschata* only produces eugenol (Supplementary Data 3). These are all phenylpropanoids synthesized from L-phenylalanine. Benzenoids are found in *R. damascena* (benzyl alcohol and benzaldehyde) and *R. gallica* (benzyl alcohol).

Two 2-phenylethanol synthesis pathways are known in rose (Supplementary Fig. 18). The first involves phenylacetaldehyde synthase gene (*PAAS*) and phenylacetaldehyde reductase gene (*PAR*)[117]. The second involves aromatic amino acid aminotransferase (AAAT3), phenylpyruvic acid decarboxylase gene (*PPDC*) and *PAR* genes[103]. We identified two *PAAS* gene copies, but only one is expressed and highly specific to open and senescent flowers (*RcHt_S1004.17*; FPKM = 30 and 11, respectively). Two *PAR* gene copies showing low constitutive expressions were identified (*RcHt_S563.20, RcHt_S1878.7*). Two gene copies of *AAAT3* exhibiting globally low expression levels were annotated (*RcHt_S60.39, RcHt_S2179.4*). Two homologous *PPDC* gene candidates (*RcHt_S356.31, RcHt_S132.46*) showed a very low expression throughout the plant, while another *PPDC* candidate (*RcHt_S334.46*) shows expression in open flowers, although at low level (Supplementary Data 3). These results are consistent with the accumulation of very small amounts of phenylpropanoid compounds, such as 2-phenylethanol, in *R. chinensis* 'Old Blush' (Supplementary Data 3).

It has been reported that eugenol biosynthesis involves the activity of the genes *PAL*, *C4H*, *4CL*, *CCoAOMT*, *CFAT*, *EGS* and *OMT1*[118]. A BLAST search using sequences from *Petunia* and basil (from Uniprot) identified two candidate gene sequences (*RcHt_240.36* and *RcHt_S589.22*) for PHENYLALANINE AMMONIA LYASE (PAL). Gene expression analyses show that these two *PAL* genes are not flower specific. Three candidates encoding for the cinnamoyl-CoA hydratase-dehydrogenase (*C4H*) were annotated as cytochrome P450 proteins. *RcHt_S14256*.1 is weakly expressed in flowers, *RcHt_S11205.1* is not expressed in open flowers but shows expression in flower buds and stamens (FPKM from 35 to 111), and *RcHt_S1491.14* shows specific expression in open flowers. They are all candidates for C4H function, although this requires to be validated by enzymatic studies. We identified four putative genes coding for putative 4-coumarate-CoA ligase (*4CL*). These genes show different expression patterns in the flower. Two among these four genes are more specific to stamens (*RcHt_S139.57* and *RcHt_S1376.17*). We identified two candidate genes coding for putative coniferyl alcohol acyltransferase (*CFAT*) (*RcHt_S292.6* and *RcHt_S1078.10*), one of which shows relatively higher expression in flower buds and in stamens. The availability of this information opens new perspectives towards the elucidating of their putative roles through enzymatic tests. The last step of eugenol biosynthesis step is catalyzed by EUGENOL SYNTHASE (EGS1). The rose homologue of EGS1 was previously characterized[119]. In *R. chinensis* 'Old blush', *RcHt_S564.16* or *RcHt_S3128.4* encodes the putative homologues of EGS1. Our expression data indicate that both genes are

expressed in 'Old Blush', thus consistent with the fact that eugenol is not produced in this rose cultivar. We identified one gene copy of the putative eugenol *O*-methyltransferase (*EOMT*) homologue, (*RcHt_S23.70*), a gene that was previously characterized in *R. chinensis* 'Spontanea'[120]. In 'Old Blush', this gene shows weak expression specific to stamens.

Benzaldehyde and benzyl alcohol biosynthesis is partially known in several plants and can be derived from t-cinnamic acid or from cinnamoyl-CoA[121]. *PAL* and *C4L* are the only known genes involved in this pathway. Homologues of these two genes were found in *R. chinensis* 'Old Blush' genome. *C4L* copies are identical to the ones identified for eugenol biosynthesis. No genes could be proposed for the last biosynthesis steps in this pathway.

To summarize, the manual annotation of genes involved in scent production allowed us to identify candidate genes in all biosynthetic pathways operating in rose flowers. Characterizing these candidate genes in other rose species with different scent characteristics will help elucidate the origin of the huge diversity of scent production in the *Rosa* genus. The rose has already been shown to synthesize some of its terpenes differently from other species, via a cytosolic nudix hydrolase. The origin and localization of the precursor of these monoterpenes, GPP, are unknown. Our study here shows that the plastidic MEP pathway genes usually involved in the GPP synthesis, have a very low expression in the flower. A more in-depth study of the contribution of the two pathways in terpenes biosynthesis in rose will show if, conversely to other plants, roses use cytosolic MVA pathway to synthesize precursors of monoterpenes.

10. Color gene pathways in rose flowers

10.1 Identification / mapping and characterization of key genes

10.1.1 Color genes

Characterized genes sequences in the flavonol / anthocyanin pathway, coming from various *Rosa* accessions (species and cultivars) were retrieved from an GenBank public database. tblastn was then used to find their closest homologs in *R. chinensis* 'Old Blush'. The genes were then mapped on the assembled haploid chromosomes. When several candidates could not be distinguished (ie. for Chalcone Synthase (CHS) or Glucosyl-Transferase 1 (GT1)) we used FPKM data (described in Supplementary Notes 9.1.2) in vegetative and floral tissues to identify the most likely candidate.

10.1.2 SPL and MYB gene families

tblastn was used to search for genes containing the conserved zinc-finger DNA binding domain characteristic of the Squamosa Promoter binding Like protein (*SPL*) gene family in the rose genome sequence. FPKM data in vegetative and floral tissues for each candidate were obtained in order to build *in-silico* expression profiles and to group *SPL* genes by functional sub-families. Particular attention was given to those SPLs that could be involved in vegetative to floral meristem transition.

Using an adapted version of WMD3 miR pipelines (Ossowski Stephan, Fitz Joffrey, Schwab Rebecca, Riester Markus and Weigel Detlef, personal communication), we build a user-friendly application facilitating the prediction of miR156 targets in the rose genome. It is based on known properties of miR/target gene interaction such as number of mismatches, no mismatch at the positions 10 and 11 (cleavage region) quality of pairing in the seed region and hybridization energy[122]. We used the canonical sequence of *Arabidopsis* miR156 (UGACAGAAGAGAGUGAGCUC) to identify its counterpart in the rose, and then we interrogated the rose genome to predict the *rose* miR156 targets.

Plant MYB proteins share a conserved R2R3 MYB domain. These transcription factors are involved in the control of cell identity and fate, cell growth and division as well as in secondary metabolism, especially the phenylpropanoid pathway. BLASTp was used to search for MYB transcription factors that have conserved R2R3 motif in the heterozygous genome. MYBs with two R2R3 motifs were kept. We retrieved 215 annotated MYB sequences for the rose. Whenever possible, the correspondence of these sequences with the homozygous annotation was established, to identify allelic copies of each MYB. Finally, 120 MYB genes corresponding to one or two allelic sequences were mapped on the homozygous pseudomolecules.

10.1.3 Real time quantitative RT-PCR

mRNA and Small RNA extraction: mRNA and small RNA were extracted from petals at three development stages (non-colored immature petals (Stage 1), petal with low anthocyanin content (Stage 2) and petal of flowers with maximum anthocyanin content (Stage 3) (Supplementary Fig. 8) using Macherey-Nagel NucleoSpin® miRNA. PVP40 was added to the samples prior to grinding. One µg RNA was treated with DNase I (Ambion® DNA-free). In order to avoid over-dilution, small RNAs were eluted on a separate column and therefore their expression had to be normalized using 5.8S rRNA. Concentration was measured using NanoDrop ND-1000 Micro-Volume (NanoDrop Technologies) before and after DNase treatment. Three biological replicates were performed for each experiment.

Small RNA quantitation: Stem-loop RT-PCR was performed as previously described (Marcial-Quino *et al.*, 2016). Reverse transcription was performed with RevertAid kit (Thermo Fisher Scientific) using primers specific to 5.8S rRNA (5.8S_RT; Supplementary Table 8) or stem-loop RT-primer for miR156 (mir156_RT, Supplementary Table 8). 5.8S rRNA and miR156 expression were quantified on QuantStudio™ 6 Flex Real-Time PCR 384 (Applied Biosystems) using Fast SYBR® Green Master Mix kit (Roche Diagnostic) using specific primers (Supplementary Table 8). Data were collected for three technical replicates per sample.

mRNA quantitation: Reverse transcription was performed using oligo-dTs (T11VN) with RevertAid kit (Thermo Fisher Scientific). The expressions of CHALCONE SYNTHASE (*CHS*), FLAVONOL SYNTHASE (*FLS*), ANTHOCYANIDIN SYNTHASE (*ANS*), FLAVONOID 3'-HYDROXYLASE (*F3'H*), DIHYDROFLAVONOL REDUCTASE and of three candidate *SPL*s (*RcHm3g0480201*, *RcHm4g0430121*, *RcHm4g0437871*) were quantified on QuantStudio™ 6 Flex Real-Time PCR 384 (Applied Biosystems) using Fast SYBR® Green Master Mix kit (Roche Diagnostic) using specific primers (Supplementary Table 9). Normalization was performed relatively to TUBULIN (*TUB*), GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE (*GAPDH*) and TRANSLATIONALLY CONTROLLED TUMOR PROTEIN (*TCTP*). Data were collected for three technical replicates per sample.

10.2 Results

The first rose cultivars arose independently in China and the peri-Mediterranean area more than 2000 years ago. Flowers of wild roses used in domestication were mostly pink or red. Breeding and selection for brightly colored flowers led to increased anthocyanin synthesis in domesticated plants when compared with their wild progenitors. Anthocyanins, in association with other polyphenolic co-pigments such as flavonols could, therefore, be considered as the main determinants of flower color diversity in cultivated roses.

Therefore, we addressed the genetic determinism and gene regulatory pathways associated with floral anthocyanins and flavonols biosynthesis that were under selection for flower color during the early history of rose cultivation and domestication.

The anthocyanin / flavonol pathway in rose flowers has been described in early 90's and most of the involved enzymes are now fully characterized. In rose flowers, the last two glycosylation steps for anthocyanin aglycone were shown to be controlled by a single glycosyl-transferase (*RhGT1*), different from other plants where these steps are achieved by the sequential action of two distinct glycosyl-transferases[123].

Although this pathway can now be considered as well described in roses, information is still lacking on how the onset of anthocyanin biosynthesis is coordinated with floral opening, which will lead to flower color variations. In *Arabidopsis thaliana*, genes controlling key steps of the anthocyanin biosynthesis, such as *DFR*, *F3'H* and *ANS*, are transcriptionally activated in stems by a MYB-bHLH-WD40 complex[124].

Over-expression of *Arabidopsis* R2R3 MYB transcription factor *AtPAP1*, leads to increased anthocyanin contents in rose petal, associated with higher emission of germacrene D[125]. This published evidence raises the possibility of a co-regulation between anthocyanin and some terpenes biosynthesis in rose flowers. *R.*

chinensis 'Old Blush' scent is composed of Germacrene D, but *PAP1* expression could not be detected during petal development. We identified that a second R2R3 MYB transcription factor, *RhMYB10,* is expressed in 'Old Blush' petals. MYB10 was previously identified and characterized as an inducer of anthocyanin biosynthesis genes in Rosaceae, including in the rose[126]. Our analyses, taken together with published data, suggest that *RhMYB10*, but not *PAP1*, acts as a common activator of anthocyanin and germacrene D synthesis (Supplementary Fig. 8; Supplementary Data 8).

10.2.1 Flavonols and anthocyanins genes in *R. chinensis* 'Old Blush'

Duplication events in first and last genes of anthocyanin biosynthesis genes.

Chalcone synthase catalyzes the condensation of malonyl-coA and coumaroyl-CoA into tetrahydroxychalcone (or naringenin chalcone), which is the initial substrate necessary for synthesizing downstream polyphenolic compounds such as flavonols and anthocyanins. We identified three genes that could potentially encode a functional CHS. Among these three genes, only one *CHSa* (*Chr1g0316441*) is expressed in 'Old Blush' flowers according to FPKM data. This gene located on chromosome 1 with two alleles, *RcHt_S637.2* and *RcHt_S2110.9*.

Other genes in the pathway were identified as single-copy, except for cyanidin 3,5-diglucosyltransferase, previously named as *RhGT1*[123]. According to our data, two functional versions of this gene stand 700 kb apart from each other on chromosome 1 (*Chr1g0378941* and *Chr1g0380121*). Only one copy (*GT1a* or *Chr1g0378941* / *RcHt_S2665.15*) is expressed in buds and opened flowers of 'Old Blush', whereas *GT1b* is expressed in vegetative organs and senescent flowers, suggesting that an initial duplication event of an ancestral glucosyl-transferase was followed in *Rosa* by a specialization of one of the two copies in order to achieve 3,5-diglucosylation of cyanidin in flowers. Orthologous genes coding for enzymes normally catalyzing the sequential two-steps glucosylation process in cyanidin mapped closely to the telomeric ends of chromosomes 1 and 2. These two genes show very low expression levels in flowers, compared to *GT1a.* Other genes in the pathway were single-copy and were mapped on *R. chinensis* 'Old Blush' pseudo-chromosomes (Figure 3).

Expression of most genes in the anthocyanin biosynthesis pathway, except F3H and GT1, increased during petal growth and pigmentation, between stage 1 and stage 3. RT-qPCR expression analyses of anthocyanin biosynthesis genes in petal (Supplementary Fig. 8b) correlated with the *in silico* expression data (Supplementary Fig. 8a). The small observed differences in expression levels could be explained by the fact that *in silico* transcriptomes were performed on bulk floral organs (sepals, stamens, carpels and hypanthium) compared to petals for the RT-qPCR experiments.

10.2.2 Regulators of anthocyanins pigments and flavonols co-pigments

Squamosa Promoter-binding Like (SPL) genes and miR156-miR157 expression patterns are consistent with a possible role in anthocyanins and flavonols synthesis. In *Arabidopsis thaliana*, anthocyanin synthesis is regulated by the miR156 - *SPL9* module in an age-dependent manner. SPL9 destabilizes the MYB-bHLH-WD40 complex, hampering anthocyanidin synthesis. High expression of miR156 promotes *SPL9*

degradation, which in turn enables anthocyanidin synthesis. In rose petals, previous report shows that miR156 expression increases in response to ethylene and negatively correlates with *SPL* expression[127]. Here, we focused on the miR156 - *SPL* regulatory module, in order to identify the transcription factors that are most likely involved in controlling anthocyanidin production in rose flowers and that could influence flower color, by its action on the formation of MYB-bHLH-WD40 complex.

Sixteen loci corresponding to putative *SPL* genes were predicted (Supplementary Fig. 9). Among them, though harboring the characteristic zinc-finger domain, *RcHt_S7297.1* is truncated. Among the 15 remaining *SPL* genes, 10 were predicted to be targets of miR156. Eight out of these 10 predicted targets show a decreased expression between floral development stage IMO (early floral organs) and OFT (open flower) (Supplementary Fig. 9). Such a decrease, although occurring in flowers instead of stems, as in *Arabidopsis*, might respond to the increase of miR156 expression during the course of floral opening. The rose gene *RcHm4g0437871* (rose *SPL9 like*) shares high identity with *AtSPL9*. We quantified rose *SPL like* expression, by RT-qPCR (Figure 3), in 'Old Blush' petals at three stages (from non-colored to maximum pigmentation at the beginning of anthesis) and then we correlated its expression with genes in the anthocyanin biosynthesis pathway. Previously, it was reported in *Arabidopsis* that accumulation of miR156 correlates with low expression of *SPL9*[124]. Our RT-qPCR quantifications of miR156 expression in rose petals show that high expression levels of miR156 correlate with a decrease of *SPL* expression during petal growth and pigmentation processes. These results, taken together with previously reported data in *Arabidopsis* and the rose, are consistent with the miR156-*SPL9* module playing a role in anthocyanin synthesis, through *SPL* destabilizing the MYB-bHLH-WD40 complex, which activates the final enzymes of the pathway in the rose (Figure 3; Supplementary Fig. 8).

Comparative RNA-seq analysis of transcriptome dynamics in *R. chinensis* showed that seven MYBs were upregulated and one MYB was down-regulation during petal growth[128]. We identified candidate MYB that show a specific pattern of expression to flower tissues at different developmental stages (Supplementary Data 8). Strikingly, only one MYB (*RcHm2g0172331*; *RcHt_S1331.19* / *RcHt_S2066.7*) was found to be highly expressed and specific to three developmental stages of the rose flower. Moreover, its expression increased from closed flower buds to open flowers. This MYB is related to At MYB21 and AtMYB24, which was previously shown to play a role in petal and stamen elongation in *Arabidopsis*[129] . MYB21 is also required for the activation of PHENYLALANINE AMMONIA-LYASE (*PAL*), the first enzyme in the phenylpropanoid pathway, that leads to secondary metabolites such as flavonoids (flavonols and anthocyanidins) and lignins (Supplementary Data 8).

RhMYB10 was previously described as an activator of *DIHYDROFLAVONOL REDUCTASE* (*DFR*), a key enzyme in the biosynthesis of anthocyanins[126]. In our functional annotation, *RhMYB10* corresponds to *RcHm3g0448721* /*RcHt_S286.29*. Its pattern of expression, mostly in closed flower buds and open flowers, is compatible with a role in the activation of anthocyanin pathway enzymes (Supplementary Data 8).

We performed phylogenetic analyses including MYB proteins from *Fragaria* and *Malus*, whose functions have been reported as activators of anthocyanin biosynthesis[130,131] (Supplementary Fig. 19). *RcHt_S286.29* from R. *chinensis* is the predicted most similar gene to rose *RhMYB10*[126], previously shown to be associated with anthocyanin biosynthesis in Rosaceae[126].

10.2.3 Coordination of pigments and volatiles synthesis

SPL genes and miR156-miR157 expression patterns are consistent with a possible role in germacrene-D synthesis

It was previously reported that over-expression of the *Arabidopsis PAP1*, a MYB activator of anthocyanin synthesis and possible sub-unit of the MYB-bHLH-WD40 transcriptional activator, in the rose triggers *ANS* overexpression but was also associated with Germacrene-D synthase (*GDS*) over-expression [125]. Two genetic copies corresponding to putative *GDS* were mapped on *R. chinensis* chromosomes. The first *GDS* gene copy, corresponding to that functionally characterized by Guterman *et al*[102], is highly expressed in the petals of opened rose flowers. The second *GDS* gene copy, is also highly expressed in petals of open flowers, but also showed high expression levels in senescing flowers (Supplementary Fig. 9). Functional characterization is needed to know if this second gene has a GDS function. Both expression patterns are evocative of the expression pattern of *ANS*. Given that *PAP1* has been suggested as a possible activator of *GDS* expression[125], we hypothesize that its action on *GDS* is mediated by the *SPL9*-miR156 regulatory module, which gives a functional basis to the necessary coordination of pigments and volatile molecule synthesis for pollinator attraction (Figure 3; Supplementary Fig. 8).

To further address this hypothesis, we compared the expression of candidates for *SPL* (*RcHm4g0437871*), *ANS* (*RcHm7g0199941*), *GDS* (*RcHm5g0038101*), and *RhMYB10* (*RcHm3g0448721*) in petals of two rose plants exhibiting contrasted flower colors: *R. chinensis* 'Sanguinea' which has petals that accumulate high levels of anthocyanins at flower opening, and *R. hybrida* 'Alister Stella Gray' which has petals that do not accumulate anthocyanins. In 'Sanguinea', *SPL* was expressed in non-colored petals (flower buds), and its expression was downregulated in colored petals (Supplementary Fig. 20), thus similar to 'Old Blush'. *SPL* expression correlated with low *ANS* and *GDS* expression in flower buds before color production (Supplementary Fig. 20). In the colored petals of 'Sanguinea', *SPL* downregulation correlated with the upregulation of both *ANS* and *GDS* expression, thus corroborating the data observed in 'Old Blush' (Figure 3b; Supplementary Fig. 20). In 'Alister Stella Gray', we observed that the expression of *SPL*, *GDS,* and *ANS* was very low at both analysed stages (flower bud and flower opening). The data show that the anti-correlation of expression between *SPL* on one side, and *ANS* and *GDS* on the other side, is observed only in the colored flower cultivars 'Sanguinea' and 'Old Blush'.

RhMYB10 exhibited similar expression patterns in both 'Sanguinea' and 'Alister Stella Gray' roses. *RhMYB10* was expressed at low levels in flower buds and its expression increased in developing petals (Supplementary Fig. 20).

The positive co-regulation of *ANS* and *GDS* expression in anthocyanins-accumulating flowers and their anti-correlated expression with *SPL* are other arguments favoring the hypothesis that anthocyanins and germacrene D biosynthesis could be coupled and achieved through the miR156-*SPL* regulatory module. These data also suggest that *RhMYB10* expression is likely not the determinant factor, but rather it is the putative action of *SPL* on MYB-bHLH-WD40 complex, which activates the final enzymes of anthocyanins and germacrene D synthesis in rose (Figure 3; Supplementary Fig. 8).

11. Auxin Response Factor gene family

Parts of this work were performed on the heterozygous assembly. The table in Supplementary Data 1 shows heterozygous IDs matched with their reference genome annotations (homozygous).

To identify Auxin Response Factor (ARF) gene family members in *R. chinensis,* the predicted proteins associated with the domain PF06507 (Auxin Response Factor) were extracted. From the 37 predicted protein sequences (Supplementary Data 5a), six were excluded from the phylogenetic analysis because they were highly truncated or contained very divergent regions (*RcHt_S12618.1, RcHt_S1403.1, RcHt_S2738.6, RcHt_S2297.1, RcHt_S2297.6,* and *RcHt_S1950.5*, indicated "No" in the column "Used for phylogenetic analyses"). These 31 protein sequences were aligned together with the sequence of 22 *Arabidopsis* ARF proteins (ARF23 was not included as it has a truncated DNA Binding Domain due to an early stop codon, and appears to be under negative selection, Supplementary Data 5b) using MAFFT (http://mafft.cbrc.jp/alignment/software/[132] with the following parameters: (1) Iterative refinement methods: G-INS-I, (2) Leave gappy regions, (3) Scoring matrix for amino acid sequences: BLOSUM62. To generate the Neighbor-Joining (NJ) tree shown in Supplementary Fig. 21, aligned protein sequences were computed with MAFFT using 198 conserved sites with the following parameters: (1) Substitution model: JTT[133], (2) Heterogeneity among sites: Estimate and (3) Boostrap resempling: 1000.

A Pfam domain search of the *Rosa chinensis* predicted protein data identified all rose representatives for the ARFs (Supplementary Fig. 21; Supplementary Data 5) except for AtARF12/13/14/15/20/21/22 clade, that has only been identified in Brassicaceae thus far. A more detailed analysis revealed that one pair of ARF sequence (*RcHt_S204.16* and *RcHt_S622.11*) have no apparent *Arabidopsis* homologs. In most cases, pairs of very closely related sequences were identified (Supplementary Fig. 21; Supplementary Data 5), underscoring the heterozygosity of *R. chinensis* genome.

12. Type II MADS-box gene family members involved in Rose flowering and flower development

Parts of this work were performed on the heterozygous assembly. The table in Supplementary Data 1 shows heterozygous IDs matched with their reference genome annotations (homozygous).

To identify type II MADS-box family members, the *R. chinensis* predicted protein dataset was searched by local BLAST analysis with BioEdit software[134], using *Arabidopsis* representatives of the major MADS-box subfamilies[135] as a template. Identified *R. chinensis* protein sequences (Supplementary Table 3) were assigned to any of the major MADS-box subfamilies based on homology scores and the presence of small conserved (C-terminal) peptide motifs that are diagnostic for the different subfamilies[136]. To generate the Neighbor-Joining (NJ) trees shown in Supplementary Fig. 22, protein sequences were first aligned using ClustalW[137] and aligned regions (Supplementary Data 6) were selected for phylogenetic analysis. NJ trees were computed with Treecon software[138] using the following parameters: (1) Distance estimation options: Tajima and Nei; Distance calculations; insertions and deletion not taken into account; Alignment positions: all; Bootstrap analysis: yes, 1000 samples. (2) Infer tree topology options: Neighbor-joining; Bootstrap analysis: yes. (3) Root unrooted trees options: outgroup option: single sequence (forced); bootstrap analysis: yes. All trees were rooted using the *Arabidopsis* FUL protein, except for the AP1/FUL subfamily, for which *Arabidopsis* SEP3 was used as an outgroup. For the phylogenetic analysis, rose and *Arabidopsis* proteins were each time compared, except for the B-function/Bsister MADS-box subfamilies, for which in addition *Petunia hybrida* representatives were included in the analysis. Some of the predicted rose MADS-box proteins mentioned in Supplementary Table 3 were excluded from the phylogenetic analysis because they were highly truncated or contained too divergent regions. These gene models may correspond to pseudo-genes or alternatively, may be due to erroneous protein predictions.

BLAST searching the *R. chinensis* predicted protein data set resulted in the identification of rose representatives (Supplementary Fig. 22; Supplementary Table 3) for all major type II MADS-box subfamilies and sublineages[135] with one notable exception (see further). In most cases, each time pairs of very closely related sequences were identified (Supplementary Fig. 22; Supplementary Table 3), underscoring the hybrid/heterozygous origin of the *R. chinensis* genome. In other cases, one of the predicted protein sequences within such a pair appeared incomplete (Supplementary Table 3; Supplementary Data 6), suggesting that these represent degenerated gene copies (pseudo-genes) or alternatively inaccurately predicted protein models. A more detailed analysis of the subfamilies encoding the floral homeotic ABC functions, show that the rose genome contains MADS-box proteins in copy numbers comparable to other eudicot species, with 1 AGL6-like gene, 3 SEP-like genes, 2 FUL-like genes, 1 AP1-like gene, 1 AP3-like gene, 1 TM6-like gene, at least 2 PI-like genes, 1 Bsister-like gene, 1 AG-like, 1 PLE-like C-function gene and 1 AGL11-like gene (D-lineage). Because rose appears to have retained a TM6-like B-function gene in parallel with its AP3-like gene, and contains more than one PI-like gene, the rose B-function more closely resembles the complex B-function of the asterid species *Petunia*[139,140] than the *Arabidopsis* B-function. Intriguingly, we failed to detect members of the flowering repressor FLC clade in rose, although *Arabidopsis* contains 6 members of this subfamily. This may suggest that FLC genes have been lost in rose, or alternatively, that rose FLC genes have diverged too strongly to be easily identified as FLC members.

13. Genetic pathways involved in diploid gamete formation

Like many crops, most rose cultivars are polyploids[141,142]. Ploidy diversity is a limiting factor in rose breeding. Most interploidy crosses lead to infertile progeny. In rose domestication, breeders have often and inefficiently attempted to tinker with ploidy levels to overcome this reproductive barrier. In order to cross wild species and tetraploid cultivars, chromosome numbers must first must be balanced: (i) Haploidization, halving the chromosome number, has been unsuccessfully attempted by *in vitro* culture of haploid cells from unfertilized ovules or ovaries and microspores or anthers. A few haploidized rose plants have been produced from *in situ* parthenogenesis induced by fertilization with pollen inactivated by irradiation. The parthenogenetic development of a haploid cell from embryo sacs into a new plant was induced and embryos were subsequently rescued by *in vitro* culture. (ii) Chromosome doubling was successfully performed by mitotic polyploidization requiring microtubule drugs to transiently block chromosome segregation in mitosis and duplicate the number of chromosomes per cell. However, *in vitro* chromosome doubling is typically associated with somaclonal variation and cytochimerism phenomenas.

The most promising alternative for rose breeders is sexual polyploidization using 2n gametes. 2n gametes were widely used in crop breeding to directly introgress new traits from diploid species into tetraploids such as *potato, manihot* or *alfalfa*[143]. They also have proved useful in recovering fertility in interspecific amphihaploid hybrids by generating new polyploids. They highly enhanced genetic diversity, heterozygosity, and heterosis. Finally, 2n gametes are very desirable as a key step in the apomictic pathway as well. In *Rosa*, 2n gamete production was demonstrated to be preponderant in hybrids, genetically controlled and dependent on environmental factors like heat[144,145]. However, to date, both environmental cues and genetic pathways giving rise to 2n gametes are too insufficiently known to be routinely used in rose breeding.

Over the last decade, genetic pathways leading to 2n gametes were identified in *Arabidopsis* and Maize. They provide a basis for developing breeding strategies that introgress new wild traits into cultivated roses and enlarge modern rose diversity and genetic background. Most orthologues of the major genes responsible for forming 2n gametes are present in the rose genome (Supplementary Fig. 23). In premeiotic pathways, endoreduplication (S6K[146]) or endomitosis (GSL8, SMT2-3[147]) events double the chromosome material before meiosis. Meiotic nuclear restitutions are the most frequent events leading to 2n gametes. They encompass different processes like meiotic cell fate specification (Argonaute)[148-150], DNA methylation [151], meiotic initiation (SWI1/DYAD[152,153]), meiosis transcriptional regulation (MMD1/DUET[154-156]), meiosis I/II transition (CYCA1;2/TAM[157-160]), OSD1[161,162], MS5/TDM1[154,163,164], SMG7[163,165], meiosis II spindle orientation (AtPS1)[162], JASON[166-168]E) and cytokinesis (MAPK signalling cascade)[156,169-175]. Disturbance in mitosis and in gametogenesis was also shown to lead to gametic genome duplication (INCENP, RBR)[176-179].

References

1. Martin, M., Piola, F., Chessel, D., Jay, M. & Heizmann, P. The domestication process of the Modern Rose: genetic structure and allelic composition of the rose complex. *Theoretical-and-Applied-Genetics* **102**, 398-404 (2001).
2. Krussmann, G. *The complete book of roses*, 436 (Timber Press xii,, Portland, 1981).
3. Smulders, M.J.M. *et al.* Genomic and Breeding Resources : Plantation and Ornamental Crops, Rose. in *Wild Crop Relatives : Genomic and Breeding Resources Plantation and Ornamental Crops* (ed. Kole, C.) (Springer-Verlag, Berlin Heidelberg, 2011).
4. Bendahmane, M., Dubois, A., Raymond, O. & Bris, M.L. Genetics and genomics of flower initiation and development in roses. *J Exp Bot* **64**, 847-57 (2013).
5. Bean, J.W. *Trees and Shrubs Hardy in the British Isles. 8th edn by D.L. Clarke and G. Taylor*, (John Murray, London, UK, 1980).
6. Phillips, R. & Rix, M. *The Quest for the Rose*, (BBC Book Publishing, London, 1993).
7. Raymond, O. Ph.D. Thesis Université Claude Bernard - Lyon1 (1999).
8. Cairns, T. *Modern roses XI, The word Encyclopidia of roses*, (Academic Press, San Diego, California, 2003).
9. Touraev, A. & Heberle-Bors, E. Microspore embryogenesis and in vitro pollen maturation in tobacco. *Methods Mol Biol* **111**, 281-91 (1999).
10. Kyo, M. & Harada, H. Control of the developmental pathway of tobacco pollen in vitro. *Planta* **168**, 427-32 (1986).
11. Heslop-Harrison, J. & Heslop-Harrison, Y. Evaluation of pollen viability by enzymatically induced fluorescence; intracellular hydrolysis of fluorescein diacetate. *Stain Technol* **45**, 115-20 (1970).
12. Vergne, P., Delvallée, I. & Dumas, C. Rapid and accurate assessment of pollen development stage in wheat and maize using DAPI and membrane permeabilization. *Stain Technol* **62**, 299-304 (1987).
13. Murashige, T. & Skoog, F. A revised medium for rapid growth and bioassays with tobacco tissue cultures. *Physiol Plant* **15**, 473-497 (1962).
14. Gamborg, O.L., Miller, R.A. & Ojima, K. Nutrient requirements of suspension cultures of soybean root cells. *Exp Cell Res* **50**, 151-8 (1968).
15. Kamo, K., Jones, B., Castillon, J., Bolar, J. & Smith, F. Dispersal and size fractionation of embryogenic callus increases the frequency of embryo maturation and conversion in hybrid tea roses. *Plant Cell Rep* **22**, 787-92 (2004).
16. Vergne, P. *et al.* Somatic embryogenesis and transformation of the diploid rose *Rosa chinensis* cv `Old Blush'. *Plant Cell Tissue and Organ Culture* **100**, 73-81 (2010).
17. Brioudes, F., Thierry, A.M., Chambrier, P., Mollereau, B. & Bendahmane, M. Translationally controlled tumor protein is a conserved mitotic growth integrator in animals and plants. *Proc Natl Acad Sci U S A* **107**, 16384-9 (2010).
18. Berlin, K. *et al.* Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol* **33**, 623-30 (2015).
19. Chin, C.S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* **13**, 1050-1054 (2016).

20. Badouin, H. *et al.* The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* **546**, 148-152 (2017).
21. VanBuren, R. *et al.* Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. *Nature* **527**, 508-11 (2015).
22. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**, 722-736 (2017).
23. Zhu, W. *et al.* Altered chromatin compaction and histone methylation drive non-additive gene expression in an interspecific Arabidopsis hybrid. *Genome Biol* **18**, 157 (2017).
24. Wang, C. *et al.* Genome-wide analysis of local chromatin packing in Arabidopsis thaliana. *Genome Res* **25**, 246-56 (2015).
25. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-93 (2009).
26. Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* **16**, 259 (2015).
27. Krueger, F. & Galore, T. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
28. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-9 (2012).
29. Akdemir, K.C. & Chin, L. HiCPlotter integrates genomic data with interaction matrices. *Genome Biol* **16**, 198 (2015).
30. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-80 (1999).
31. Huang, X. & Madan, A. CAP3: A DNA sequence assembly program. *Genome Res* **9**, 868-77 (1999).
32. Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-2 (2015).
33. Dubois, A. *et al.* Transcriptome database resource and gene expression atlas for the rose. *BMC Genomics* **13**, 638-648 (2012).
34. Mott, R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput Appl Biosci* **13**, 477-8 (1997).
35. Dubois, A. *et al.* Tinkering with the C-function: a molecular frame for the selection of double flowers in cultivated roses. *PLoS One* **5**, e9288 (2010).
36. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).
37. Stajich, J.E. *et al.* The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* **12**, 1611-8 (2002).
38. Yan, H. *et al.* The Rosa chinensis cv. Viridiflora Phyllody Phenotype Is Associated with Misexpression of Flower Organ Identity Genes. *Front Plant Sci* **7**, 996 (2016).
39. Foissac S, G.J., Rombauts S, Mathe C, Amselem J, Sterck L, Van de Peer Y, Rouze P, Schiex T Genome annotation in plants and fungi: EuGene as a model platform. . *Current Bioinformatics* **3**, 87–97 (2008).
40. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
41. Lamesch, P. *et al.* The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res* **40**, D1202-10 (2012).
42. International Brachypodium, I. Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature* **463**, 763-8 (2010).

43. Bao, W., Kojima, K.K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**, 11 (2015).
44. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-9 (2008).
45. Koning-Boucoiran, C.F. *et al.* Using RNA-Seq to assemble a rose transcriptome with more than 13,000 full-length expressed genes and to develop the WagRhSNP 68k Axiom SNP array for rose (Rosa L.). *Front Plant Sci* **6**, 249 (2015).
46. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-75 (2005).
47. Girgis, H.Z. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinformatics* **16**, 227 (2015).
48. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-64 (1997).
49. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35**, 3100-8 (2007).
50. Nawrocki, E.P. *et al.* Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res* **43**, D130-7 (2015).
51. Shulaev, V. *et al.* The genome of woodland strawberry (Fragaria vesca). *Nat Genet* **43**, 109-16 (2011).
52. VanBuren, R. *et al.* The genome of black raspberry (Rubus occidentalis). *Plant J* **87**, 535-47 (2016).
53. Velasco, R. *et al.* The genome of the domesticated apple (Malus x domestica Borkh.). *Nat Genet* **42**, 833-9 (2010).
54. Daccord, N. *et al.* High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nat Genet* **49**, 1099-1106 (2017).
55. Chagne, D. *et al.* The draft genome sequence of European pear (Pyrus communis L. 'Bartlett'). *PLoS One* **9**, e92644 (2014).
56. Wu, J. *et al.* The genome of the pear (Pyrus bretschneideri Rehd.). *Genome Res* **23**, 396-408 (2013).
57. Zhang, Q. *et al.* The genome of Prunus mume. *Nat Commun* **3**, 1318 (2012).
58. International Peach Genome, I. *et al.* The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat Genet* **45**, 487-94 (2013).
59. Li, Y. *et al.* De novo assembly and characterization of the fruit transcriptome of Chinese jujube (Ziziphus jujuba Mill.) Using 454 pyrosequencing and the development of novel tri-nucleotide SSR markers. *PLoS One* **9**, e106438 (2014).
60. Huang, J. *et al.* The Jujube Genome Provides Insights into Genome Evolution and the Domestication of Sweetness/Acidity Taste in Fruit Trees. *PLoS Genet* **12**, e1006433 (2016).
61. Krishnakumar, V. *et al.* MTGD: The Medicago truncatula genome database. *Plant Cell Physiol* **56**, e1 (2015).
62. Young, N.D. *et al.* The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature* **480**, 520-4 (2011).
63. Martinez-Garcia, P.J. *et al.* The walnut (Juglans regia) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols. *Plant J* **87**, 507-32 (2016).
64. Tuskan, G.A. *et al.* The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* **313**, 1596-604 (2006).

65. Ming, R. *et al.* The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature* **452**, 991-6 (2008).
66. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463-7 (2007).
67. Tomato Genome, C. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635-41 (2012).
68. Goff, S.A. *et al.* A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science* **296**, 92-100 (2002).
69. Veluchamy, A. *et al.* LHP1 Regulates H3K27me3 Spreading and Shapes the Three-Dimensional Conformation of the Arabidopsis Genome. *PLoS One* **11**, e0158936 (2016).
70. Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data. *Available at:* http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
71. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
72. Wysoker, A., Tibbetts, K. & Fennell, T. Picard tools version 1.90. http://picard.sourceforge.net/. (2013).
73. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
74. Zang, C. *et al.* A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **25**, 1952-8 (2009).
75. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-89 (2010).
76. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-2 (2010).
77. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-45 (2009).
78. Shen, L., Shao, N., Liu, X. & Nestler, E. ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics* **15**, 284 (2014).
79. Ye, T. *et al.* seqMINER: an integrated ChIP-seq data interpretation platform. *Nucleic Acids Res* **39**, e35 (2011).
80. Bannister, A.J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res* **21**, 381-95 (2011).
81. Lauria, M. & Rossi, V. Epigenetic control of gene regulation in plants. *Biochim Biophys Acta* **1809**, 369-78 (2011).
82. Rodriguez-Granados, N.Y. *et al.* Put your 3D glasses on: plant chromatin is on show. *Journal of Experimental Botany* **67**, 3205–3221 (2016).
83. Shi, J. & Dawe, R.K. Partitioning of the maize epigenome by the number of methyl groups on histone H3 lysines 9 and 27. *Genetics* **173**, 1571-83 (2006).
84. Slotkin, R.K. Plant epigenetics: from genotype to phenotype and back again. *Genome Biol* **17**, 57 (2016).
85. Latrasse, D. *et al.* The quest for epigenetic regulation underlying unisexual flower development in Cucumis melo. *Epigenetics Chromatin* **10**, 22 (2017).
86. Salse, J. Ancestors of modern plant crops. *Curr Opin Plant Biol* **30**, 134-42 (2016).
87. Longhi, S. *et al.* Molecular genetics and genomics of the Rosoideae: state of the art and future perspectives. *Hortic Res* **1**, 1 (2014).

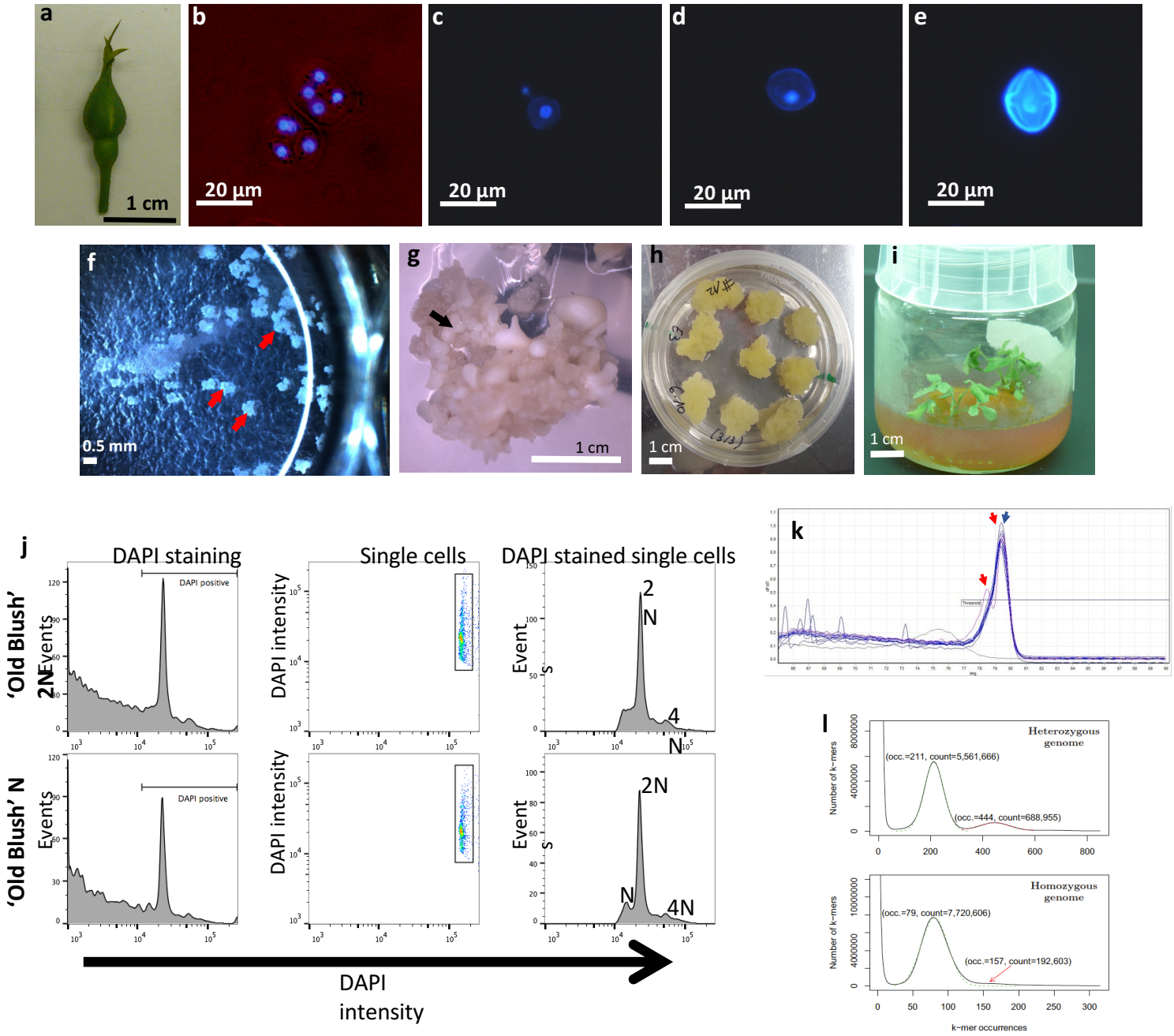
88. Xiang, Y. *et al.* Evolution of Rosaceae Fruit Types Based on Nuclear Phylogeny in the Context of Geological Times and Genome Duplication. *Mol Biol Evol* **34**, 262-281 (2017).
89. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-7 (2004).
90. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**, 540-52 (2000).
91. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764-70 (2011).
92. Koboldt, D.C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283-5 (2009).
93. Redher, A. *Manual of cultivated trees and shrubs, hardy in North America*, (New York., 1940).
94. Zhu, Z.M., Gao, X.F. & Fougere-Danezan, M. Phylogeny of Rosa sections Chinenses and Synstylae (Rosaceae) based on chloroplast and nuclear markers. *Mol Phylogenet Evol* **87**, 50-64 (2015).
95. Bergougnoux, V. *et al.* Both the adaxial and abaxial epidermal layers of the rose petal emit volatile scent compounds. *Planta* **226**, 853-66 (2007).
96. Adams, R.P. *Identification of Essential Oil Components By Gas Chromatography/Mass Spectrometry, 4th Edition*, (Allured Publishing Corporation, Carol Stream, Illinois, USA, 2007).
97. Picone, J.M., Clery, R.A., Watanabe, N., MacTavish, H.S. & Turnbull, C.G. Rhythmic emission of floral volatiles from Rosa damascena semperflorens cv. 'Quatre Saisons'. *Planta* **219**, 468-78 (2004).
98. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).
99. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27**, 2325-9 (2011).
100. Carver, T., Harris, S.R., Berriman, M., Parkhill, J. & McQuillan, J.A. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **28**, 464-469 (2012).
101. Chen, X.M. *et al.* Functional characterization of rose phenylacetaldehyde reductase (PAR), an enzyme involved in the biosynthesis of the scent compound 2-phenylethanol. *Journal of Plant Physiology* **168**, 88-95 (2011).
102. Guterman, I. *et al.* Rose scent: genomics approach to discovering novel floral fragrance-related genes. *Plant Cell* **14**, 2325-38 (2002).
103. Hirata, H. *et al.* Seasonal induction of alternative principal pathway for rose flower scent. *Scientific Reports* **6**, 20234 (2016).
104. Magnard, J.L. *et al.* PLANT VOLATILES. Biosynthesis of monoterpene scent compounds in roses. *Science* **349**, 81-3 (2015).
105. Scalliet, G. *et al.* Scent evolution in Chinese roses. *Proc Natl Acad Sci U S A* **105**, 5927-32 (2008).
106. Scalliet, G. *et al.* Role of petal-specific orcinol O-methyltransferases in the evolution of rose scent. *Plant Physiol* **140**, 18-29 (2006).
107. Wu, S. *et al.* The key role of phloroglucinol O-methyltransferase in the biosynthesis of Rosa chinensis volatile 1,3,5-trimethoxybenzene. *Plant Physiol* **135**, 95-102 (2004).

108. Meslet-Cladiere, L. *et al.* Structure/function analysis of a type iii polyketide synthase in the brown alga Ectocarpus siliculosus reveals a biochemical pathway in phlorotannin monomer biosynthesis. *Plant Cell* **25**, 3089-103 (2013).
109. Taura, F. *et al.* A Novel Class of Plant Type III Polyketide Synthase Involved in Orsellinic Acid Biosynthesis from Rhododendron dauricum. *Front Plant Sci* **7**, 1452 (2016).
110. Hemmerlin, A., Harwood, J.L. & Bach, T.J. A raison d'etre for two distinct pathways in the early steps of plant isoprenoid biosynthesis? *Prog Lipid Res* **51**, 95-148 (2012).
111. Sun, P., Schuurink, R.C., Caissard, J.C., Hugueney, P. & Baudino, S. My Way: Noncanonical Biosynthesis Pathways for Plant Volatiles. *Trends Plant Sci* **21**, 884-894 (2016).
112. Huang, F.C. *et al.* Substrate promiscuity of RdCCD1, a carotenoid cleavage oxygenase from Rosa damascena. *Phytochemistry* **70**, 457-64 (2009).
113. Feng, L. *et al.* Flowery odor formation revealed by differential expression of monoterpene biosynthetic genes and monoterpene accumulation in rose (Rosa rugosa Thunb.). *Plant Physiol Biochem* **75**, 80-8 (2014).
114. Degenhardt, J., Kollner, T.G. & Gershenzon, J. Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. *Phytochemistry* **70**, 1621-37 (2009).
115. Scala, A., Allmann, S., Mirabella, R., Haring, M.A. & Schuurink, R.C. Green Leaf Volatiles: A Plant's Multifunctional Weapon against Herbivores and Pathogens. *International Journal of Molecular Sciences* **14**, 17781-17811 (2013).
116. Shalit, M. *et al.* Volatile ester formation in roses. Identification of an acetyl-coenzyme A. Geraniol/Citronellol acetyltransferase in developing rose petals. *Plant Physiol* **131**, 1868-76 (2003).
117. Kaminaga, Y. *et al.* Plant phenylacetaldehyde synthase is a bifunctional homotetrameric enzyme that catalyzes phenylalanine decarboxylation and oxidation. *J Biol Chem* **281**, 23357-66 (2006).
118. Koeduka, T. The phenylpropene synthase pathway and its applications in the engineering of volatile phenylpropanoids in plants. *Plant Biotechnology* **31**, 401-407 (2014).
119. Yan, H. *et al.* Cloning and expression analysis of eugenol synthase gene (RhEGS1) in cut rose (Rosa hybrida). *Scientia Agricultura sinica* **45**, 590-597 (2012).
120. Wu, S. *et al.* Two O-methyltransferases isolated from flower petals of Rosa chinensis var. spontanea involved in scent biosynthesis. *J Biosci Bioeng* **96**, 119-28 (2003).
121. Widhalm, J.R. & Dudareva, N. A familiar ring to it: biosynthesis of plant benzoic acids. *Mol Plant* **8**, 83-97 (2015).
122. Schwab, R. *et al.* Specific Effects of MicroRNAs on the Plant Transcriptome. *Developmental Cell* **8**, 517-527 (2005).
123. Ogata, J., Kanno, Y., Itoh, Y., Tsugawa, H. & Suzuki, M. Plant biochemistry: anthocyanin biosynthesis in roses. *Nature* **435**, 757-8 (2005).
124. Gou, J.Y., Felippes, F.F., Liu, C.J., Weigel, D. & Wang, J.W. Negative regulation of anthocyanin biosynthesis in Arabidopsis by a miR156-targeted SPL transcription factor. *Plant Cell* **23**, 1512-22 (2011).
125. Zvi, M.M. *et al.* PAP1 transcription factor enhances production of phenylpropanoid and terpenoid scent compounds in rose flowers. *New Phytol* **195**, 335-45 (2012).
126. Lin-Wang, K. *et al.* An R2R3 MYB transcription factor associated with regulation of the anthocyanin biosynthetic pathway in Rosaceae. *BMC Plant Biol* **10**, 50 (2010).
127. Pei, H. *et al.* Integrative analysis of miRNA and mRNA profiles in response to ethylene in rose petals during flower opening. *PLoS One* **8**, e64290 (2013).

128. Han, Y. *et al.* Comparative RNA-seq analysis of transcriptome dynamics during petal development in Rosa chinensis. *Sci Rep* **7**, 43382 (2017).
129. Reeves, P.H. *et al.* A regulatory network for coordinated flower maturation. *PLoS Genet* **8**, e1002506 (2012).
130. An, X.H. *et al.* MdMYB9 and MdMYB11 are involved in the regulation of the JA-induced biosynthesis of anthocyanin and proanthocyanidin in apples. *Plant Cell Physiol* **56**, 650-62 (2015).
131. Schaart, J.G. *et al.* Identification and characterization of MYB-bHLH-WD40 regulatory complexes controlling proanthocyanidin biosynthesis in strawberry (Fragaria x ananassa) fruits. *New Phytol* **197**, 454-67 (2013).
132. Katoh, K. & Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-80 (2013).
133. Jones, D.T., Taylor, W.R. & Thornton, J.M. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**, 275-82 (1992).
134. Hall, T.A. BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. . *Nucleic Acids Symp. Ser.* **41**, 95–98. (1999).
135. Becker, A. & Theissen, G. The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Mol Phylogenet Evol* **29**, 464-89 (2003).
136. Vandenbussche, M. *et al.* Toward the analysis of the petunia MADS box gene family by reverse and forward transposon insertion mutagenesis approaches: B, C, and D floral organ identity functions require SEPALLATA-like MADS box genes in petunia. *Plant Cell* **15**, 2680-93 (2003).
137. Thompson, J.D., Higgins, D.G. & Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-80 (1994).
138. Van de Peer, Y. & De Wachter, R. TREECON for Windows: S software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. . *Comput. Appl. Biosci.* **10**, 569–570. (1994).
139. Rijpkema, A.S. *et al.* Analysis of the Petunia TM6 MADS box gene reveals functional divergence within the DEF/AP3 lineage. *Plant Cell* **18**, 1819-32 (2006).
140. Vandenbussche, M., Zethof, J., Royaert, S., Weterings, K. & Gerats, T. The duplicated B-class heterodimer model: whorl-specific effects and complex genetic interactions in Petunia hybrida flower development. *Plant Cell* **16**, 741-54 (2004).
141. Bennett, M.D. Perspectives on polyploidy in plants – ancient and neo. *Biological Journal of the Linnean Society* **82**, 411–423 (2004).
142. Udall, J.A. & Wendel, J.F. Polyploidy and Crop Improvement. *Crop Science* **46**, S–3 (2006).
143. Brownfield, L. & Köhler, C. Unreduced gamete formation in plants: mechanisms and prospects. *Journal of Experimental Botany* **62**, 1659–1668 (2011).
144. Crespel, L., Ricci, S.C. & Gudin, S. The production of 2n pollen in rose. *Euphytica* **151**, 155–164 (2006).
145. Pécrix, Y. *et al.* Polyploidization mechanisms: temperature environment can induce diploid gamete formation in Rosa sp. *Journal of Experimental Botany* **62**, 3587–3597 (2011).
146. Henriques, R. *et al.* Arabidopsis S6 kinase mutants display chromosome instability and altered RBR1–E2F pathway activity. *The EMBO Journal* **29**, 2979–2993 (2010).

147. De Storme, N. *et al.* GLUCAN SYNTHASE-LIKE8 and STEROL METHYLTRANSFERASE2 Are Required for Ploidy Consistency of the Sexual Reproduction System in Arabidopsis. *The Plant Cell* **25**, 387–403 (2013).
148. Hernandez-Lagana, E., Rodriguez-Leal, D., Lua, J. & Vielle-Calzada, J.-P. A Multigenic Network of ARGONAUTE4 Clade Members Controls Early Megaspore Formation in Arabidopsis. *Genetics* **204**, 1045–1056 (2016).
149. Olmedo-Monfil, V. *et al.* Control of female gamete formation by a small RNA pathway in Arabidopsis. *Nature* **464**, 628–632 (2010).
150. Singh, M. *et al.* Production of Viable Gametes without Meiosis in Maize Deficient for an ARGONAUTE Protein. *The Plant Cell* **23**, 443–458 (2011).
151. Garcia-Aguilar, M., Michaud, C., Leblanc, O. & Grimanelli, D. Inactivation of a DNA Methylation Pathway in Maize Reproductive Organs Results in Apomixis-Like Phenotypes. *The Plant Cell* **22**, 3249–3267 (2010).
152. Mercier, R. *et al.* SWITCH1 (SWI1): a novel protein required for the establishment of sister chromatid cohesion and for bivalent formation at meiosis. *Genes & Development* **15**, 1859–1871 (2001).
153. Pawlowski, W.P. *et al.* Maize AMEIOTIC1 is essential for multiple early meiotic processes and likely required for the initiation of meiosis. *Proceedings of the National Academy of Sciences* **106**, 3603–3608 (2009).
154. Andreuzza, S., Nishal, B., Singh, A. & Siddiqi, I. The Chromatin Protein DUET/MMD1 Controls Expression of the Meiotic Gene TDM1 during Male Meiosis in Arabidopsis. *PLoS Genet* **11**, e1005396 (2015).
155. Reddy, T.V., Kaur, J., Agashe, B., Sundaresan, V. & Siddiqi, I. The DUET gene is necessary for chromosome organization and progression during male meiosis in Arabidopsis and encodes a PHD finger protein. *Development* **130**, 5975–5987 (2003).
156. Yang, C.-Y. *et al.* TETRASPORE encodes a kinesin required for male meiotic cytokinesis in Arabidopsis. *The Plant Journal* **34**, 229–240 (2003).
157. d'Erfurth, I. *et al.* The CYCLIN-A CYCA1;2/TAM Is Required for the Meiosis I to Meiosis II Transition and Cooperates with OSD1 for the Prophase to First Meiotic Division Transition. *PLoS Genetics* **6**(2010).
158. Magnard, J.-L., Yang, M., Chen, Y.-C.S., Leary, M. & McCormick, S. The Arabidopsis Gene Tardy Asynchronous Meiosis Is Required for the Normal Pace and Synchrony of Cell Division during Male Meiosis. *Plant Physiology* **127**, 1157–1166 (2001).
159. Wang, Y., Jha, A.K., Chen, R., Doonan, J.H. & Yang, M. Polyploidy-associated genomic instability in Arabidopsis thaliana. *genesis* **48**, 254–263 (2010).
160. Wang, Y., Magnard, J.-L., McCormick, S. & Yang, M. Progression through Meiosis I and Meiosis II in Arabidopsis Anthers Is Regulated by an A-Type Cyclin Predominately Expressed in Prophase I. *Plant Physiology* **136**, 4127–4135 (2004).
161. Cromer, L. *et al.* OSD1 Promotes Meiotic Progression via APC/C Inhibition and Forms a Regulatory Network with TDM and CYCA1;2/TAM. *PLoS Genet* **8**, e1002865 (2012).
162. d'Erfurth, I. *et al.* Turning Meiosis into Mitosis. *PLoS Biology* **7**(2009).
163. Bulankova, P., Riehs-Kearnan, N., Nowack, M.K., Schnittger, A. & Riha, K. Meiotic Progression in Arabidopsis Is Governed by Complex Regulatory Interactions between SMG7, TDM1, and the Meiosis I–Specific Cyclin TAM. *The Plant Cell* **22**, 3791–3803 (2010).
164. Cifuentes, M. *et al.* TDM1 Regulation Determines the Number of Meiotic Divisions. *PLOS Genetics* **12**, e1005856 (2016).

165. Riehs, N. *et al.* Arabidopsis SMG7 protein is required for exit from meiosis. *Journal of Cell Science* **121**, 2208–2216 (2008).
166. Brownfield, L. *et al.* Organelles maintain spindle position in plant meiosis. *Nature Communications* **6**(2015).
167. De Storme, N. & Geelen, D. The Arabidopsis Mutant jason Produces Unreduced First Division Restitution Male Gametes through a Parallel/Fused Spindle Mechanism in Meiosis II. *Plant Physiology* **155**, 1403–1415 (2011).
168. Erilova, A. *et al.* Imprinting of the Polycomb Group Gene MEDEA Serves as a Ploidy Sensor in Arabidopsis. *PLoS Genet* **5**, e1000663 (2009).
169. Hülskamp, M. *et al.* The STUD Gene Is Required for Male-Specific Cytokinesis after Telophase II of Meiosis in Arabidopsis thaliana. *Developmental Biology* **187**, 114–124 (1997).
170. Oh, S.-A., Bourdon, V., Das'Pal, M., Dickinson, H. & Twell, D. Arabidopsis Kinesins HINKEL and TETRASPORE Act Redundantly to Control Cell Plate Expansion during Cytokinesis in the Male Gametophyte. *Molecular Plant* **1**, 794–799 (2008).
171. Sasabe, M. *et al.* Phosphorylation of a mitotic kinesin-like protein and a MAPKKK by cyclin-dependent kinases (CDKs) is involved in the transition to cytokinesis in plants. *Proceedings of the National Academy of Sciences* **108**, 17844–17849 (2011).
172. Spielman, M. *et al.* TETRASPORE is required for male meiotic cytokinesis in Arabidopsis thaliana. *Development* **124**, 2645–2657 (1997).
173. Takahashi, Y., Soyano, T., Kosetsu, K., Sasabe, M. & Machida, Y. HINKEL kinesin, ANP MAPKKKs and MKK6/ANQ MAPKK, which phosphorylates and activates MPK4 MAPK, constitute a pathway that is required for cytokinesis in Arabidopsis thaliana. *Plant and Cell Physiology* **51**, 1766–1776 (2010).
174. Tanaka, H. *et al.* The AtNACK1/HINKEL and STUD/TETRASPORE/AtNACK2 genes, which encode functionally redundant kinesins, are essential for cytokinesis in Arabidopsis. *Genes to Cells* **9**, 1199–1211 (2004).
175. Zeng, Q., Chen, J.-G. & Ellis, B.E. AtMPK4 is required for male-specific meiotic cytokinesis in Arabidopsis. *The Plant Journal* **67**, 895–906 (2011).
176. Ebel, C., Mariconti, L. & Gruissem, W. Plant retinoblastoma homologues control nuclear proliferation in the female gametophyte. *Nature* **429**, 776–780 (2004).
177. Johnston, A.J. *et al.* Dosage-Sensitive Function of RETINOBLASTOMA RELATED and Convergent Epigenetic Control Are Required during the Arabidopsis Life Cycle. *PLOS Genetics* **6**, e1000988 (2010).
178. Johnston, A.J., Matveeva, E., Kirioukhova, O., Grossniklaus, U. & Gruissem, W. A Dynamic Reciprocal RBR-PRC2 Regulatory Circuit Controls Arabidopsis Gametophyte Development. *Current Biology* **18**, 1680–1686 (2008).
179. Kirioukhova, O. *et al.* Female gametophytic cell specification and seed development require the function of the putative Arabidopsis INCENP ortholog WYRD. *Development* **138**, 3409–3420 (2011).



Supplementary Figure 1. Extraction of homozygous material from heterozygous R. *chinensis* 'Old Blush' by *in vitro* microspore culture.

a, Floral bud when most microspores are at the mid-late uninucleate/early bicellular development stages.

b-e, DAPI staining of microspores developmental stages. (b) two tetrads, (c) early uninucleate, (d) mid-uninucleate, (e) mid-bicellular pollen grain with autofluorescent wall.

f-h, Identification (**f,** red arrows) and multiplication of homozygous microcali obtained from microspores culture.

g, callus with somatic embryos (arrow). **h,** multiplication of *Rc*HzRDP12 homozygous calli..

i, Plantlet regenerated from *Rc*HzRDP12 homozygous callus.

j, Fluorescence-activated cell sorting analysis shows that the obtained homozygous *Rc*HzRDP12 underwent spontaneous genome duplication during regeneration resulting in diploid homozygous callus with a similar ploidy profile as the heterozygous *R. chinensis* 'Old Blush' plants.

k, HRM analyses to amplify heterozygous loci in 'Old Blush' genome. Red arrows indicate the heterozygous loci in 'Old Blush' genome. All tested loci (blue arrow) showed that the *Rc*HzRDP12 genome was homozygous.

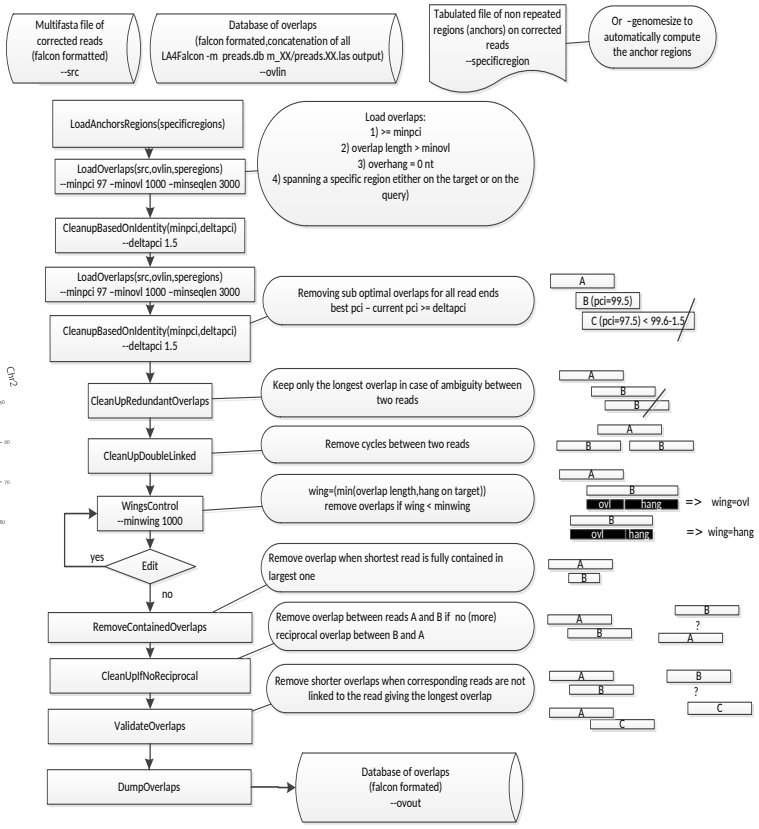
l, Compared *k*-mer frequency distribution in heterozygous and homozygous *Rosa chinensis* genomes. *k*-mers of length 47 were counted using Jellyfish[18] in the whole raw Illumina datasets and the number of distinct *k*-mers was plotted against their number of occurrences in the reads. The top plot displays two peaks, at 211 and 444, denoting the existence of two types of regions in the genome: some present in one copy (occ.=211), and some present in two copies (occ.=444 ≈2*211), consistent with the hypothesis that most of the genome is highly heterozygous (one copy), while a smaller part is homozygous (two copies). In the homozygous genome (bottom plot), only one peak remains, confirming that we extracted one single haplotype from *R. chinensis* 'Old Blush' heterozygous genome; a very small bump can be seen on the right (occ.=157), which could correspond to tandem duplications in the extracted haplotype.

a)

Step	Software	Main parameters	# of sequences	Min length (BP)	Max length (BP)	N50 (BP)	# contigs larger than N50	MEAN (BP)	MEDIAN (BP)	Assembly size (BP)	# of Ns
1	assembly of mitochondrial and chloroplast genomes	1 SMRT Cell (1.7Gb of subreads)		CP	MT						
	CANU 1.3	genomeSize=560m -pacbio-raw, minlen 50000	2	183 433	313 027					496 460	0
2	Contig assemblies	40 SMRT Cells									
	CANU 1.4	genomeSize=560m -pacbio-raw corOutCoverage=100	393	7 267	24 778 677	6 981 035	22	1 321 229	95 253	519 243 116	0
	CANU 1.4	genomeSize=560m -pacbio-raw [corOutCoverage=40]	413	1 670	21 312 326	7 955 998	22	1 252 030	74 181	517 088 455	0
	Falcon/til-r-20161228	--minpci 97 --deltapci 1.5 [--minovl 1000 --minwing 1000]	322	1 203	14 017 564	4 936 831	36	1 564 588	388 732	503 797 445	0
	Falcon/til-r-20161228	--minpci 97 --deltapci 1.5 --minovl 2000 --minwing 4000	296	1 426	14 017 521	5 084 517	36	1 691 472	486 918	500 675 824	0
	Falcon/til-r-20161228	--minpci 97.5 --deltapci 2 --minovl 2000 --minwing 4000	298	1 426	15 813 146	4 747 884	35	1 660 874	428 231	494 940 690	0
	Falcon/til-r-20161228	--minpci 98 --deltapci 2 --minovl 2000 --minwing 4000	298	6 751	13 902 803	3 373 044	44	1 614 034	822 346	480 982 429	0
3	Meta assembly	2 CANU + 4 FALCON assemblies									
	CANU 1.4	cnsConsensus=utgcns minOverlapLength=10000 minReadLength=10000	82	69 763	53 182 455	24 335 301	7	6 280 144	302 069	514 971 815	0
4	Polishing	Meta assembly (step 3) + mitochondrial and chloroplast genomes (step 1)									
	blasr/quiver	**blasr:** --minLength 3000 --maxHits 1	84	69 832	53 183 171	24 340 895	7	6 136 904	302 550	515 499 949	0
	glint/samtools/pilon	**glint:** --best-score --mmis 10 --lrmin 0.8 --lmin 80 --step 2 --no-lc-filtering; **samtools:** -f 0x02 -q 10; **pilon:** --mindepth 30 --fix bases	84	69 833	53 194 914	24 346 855	7	6 138 247	302 564	515 612 804	0
5	Genetic map integration	Identification and breaking 4 validated breakpoints									
	ALLMAPS, bedtools maskseq	**blast :** -evalue 10e-18 -perc_identity 95 **inhouse_script_for_filtering** = hsp_length > 56, unique_match=1 **allmaps:** chunk=2, chunk=4	88	69 833	42 739 042	22 201 688	9	5 859 236	318 657	515 612 796	0
6	Pseudomolecule building		7 chr, MT, CP, 46 on chr0	# of scaffolds							
	ALLMAPS + chloroplast circularization	100 N per gap	55	69 833	89 953 796	69 643 165	4	9 374 344	199 463	515 588 973	3300

b)

c)

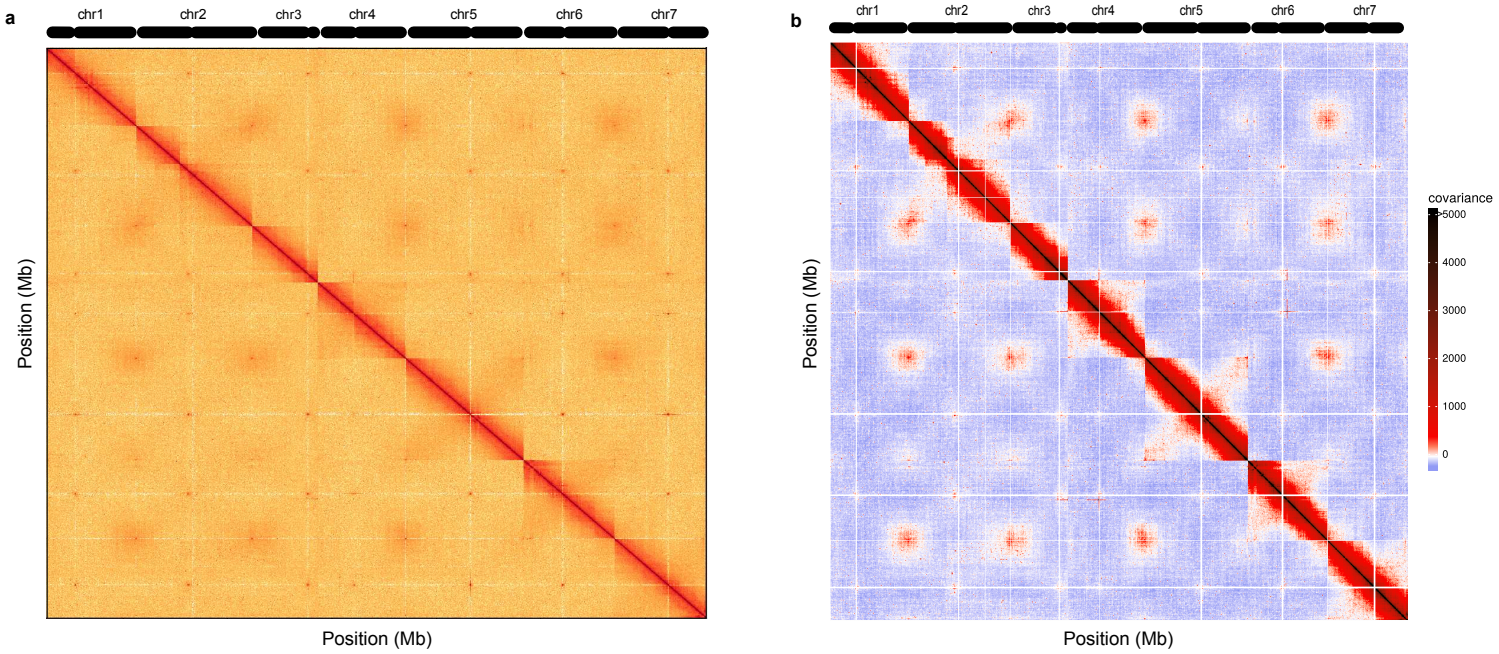


Supplementary Figure 2. Meta-assembly of the *Rosa chinensis* 'Old Blush' genome.

a, Summary of the assembly process including software, version and parameters, and the evolution of the assembly statistics during the process**.**

b, Visualisation of gaps in CANU and FALCON primary assemblies. a-f) regions absent in primary assemblies obtained with CANU and FALCON/til-r are coloured in blue. a, CANU release 1.4, default parameters. b, CANU version 1.4, corOutCoverage=100. c, Falcon/til-r, minimum length of the overlap=1000nt (minovl), minimum percentage of identity=97 (minpci), maximum difference of identity percentage=1.5 (deltapci), minimum dangling length=1000nt (minwing). d, Falcon/til-r, minpci=97, deltapci=1.5, minovl=2000, minwing=4000. e, Falcon/til-r, minpci=97.5, deltapci=2, minovl=2000, minwing=4000. f, Falcon/til-r, minpci=98, deltapci=2, minovl=2000, minwing=4000. g, regions that are absent in the six primary assemblies (a-f) are coloured in blue. h, regions corresponding to nucleotide gaps in the pseudomolecules (stretches of N) are represented in black. i, mean coverage obtained by mapping Pacbio corrected reads, window size = 250kb.

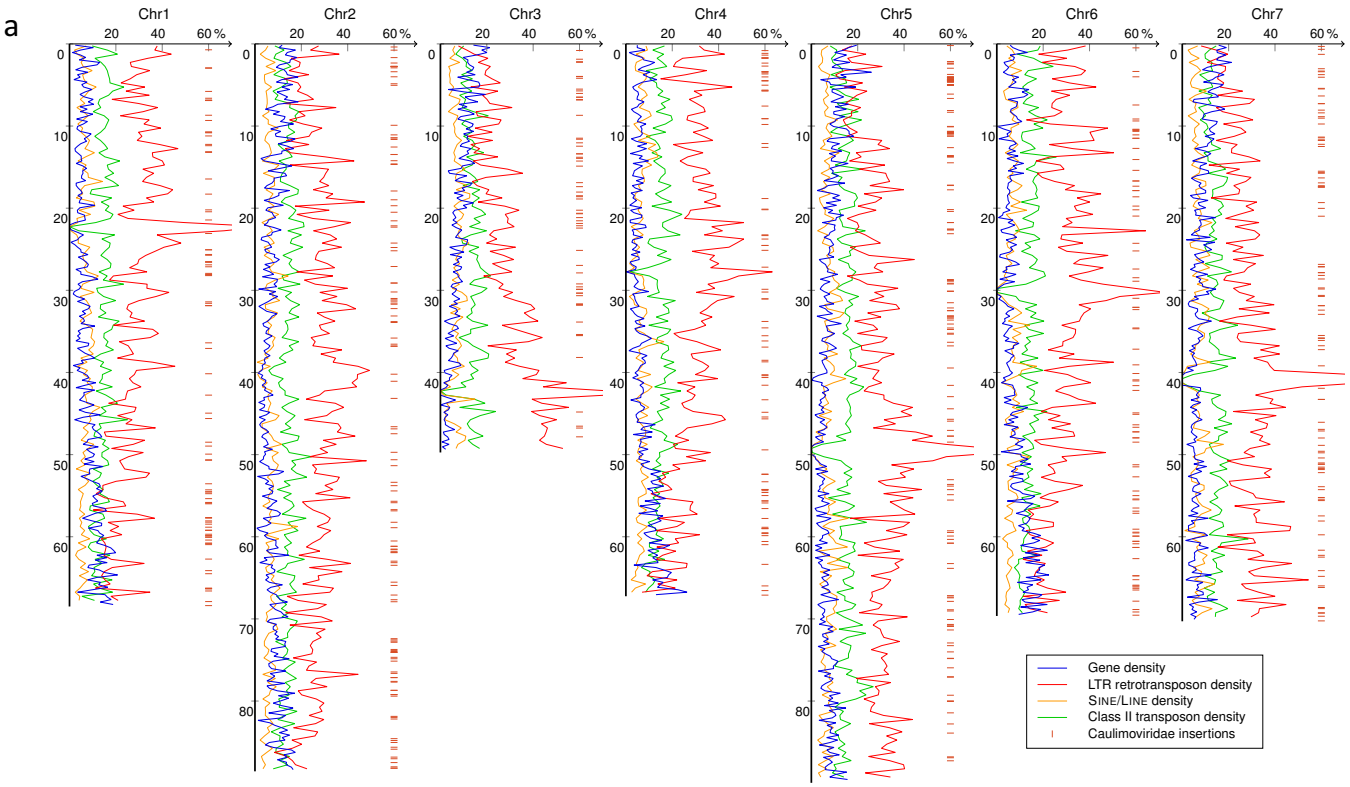
c, Overview of the different modules of the til-r software.

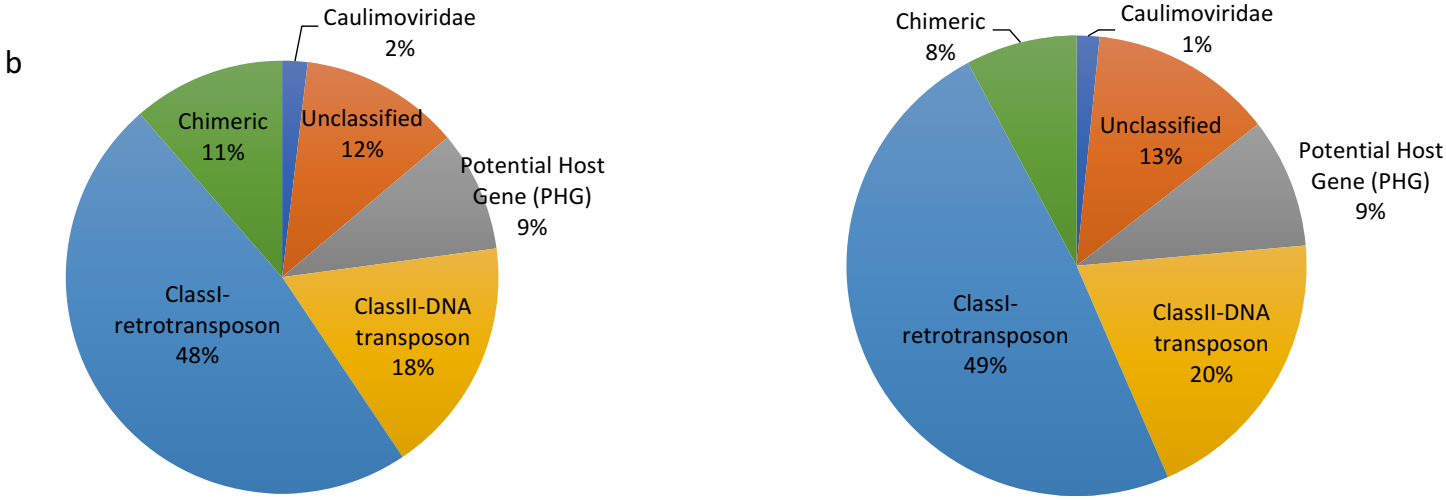


Supplementary Figure 3. Chromosomal Hi-C contact map data analysis.

a, Inter-chromosomal Hi-C contact map. The intensity of each pixel represents the count of Hi-C links between 400kb windows on chromosomes on a logarithmic scale. Darker red pixels indicate a higher contact probabilities.

b, Covariance matrix. Each dot represents the covariance between two values *i* and *j* on *x*-axis and *y*-axis, respectively. Each value is the number of interactions observed every 500kb.





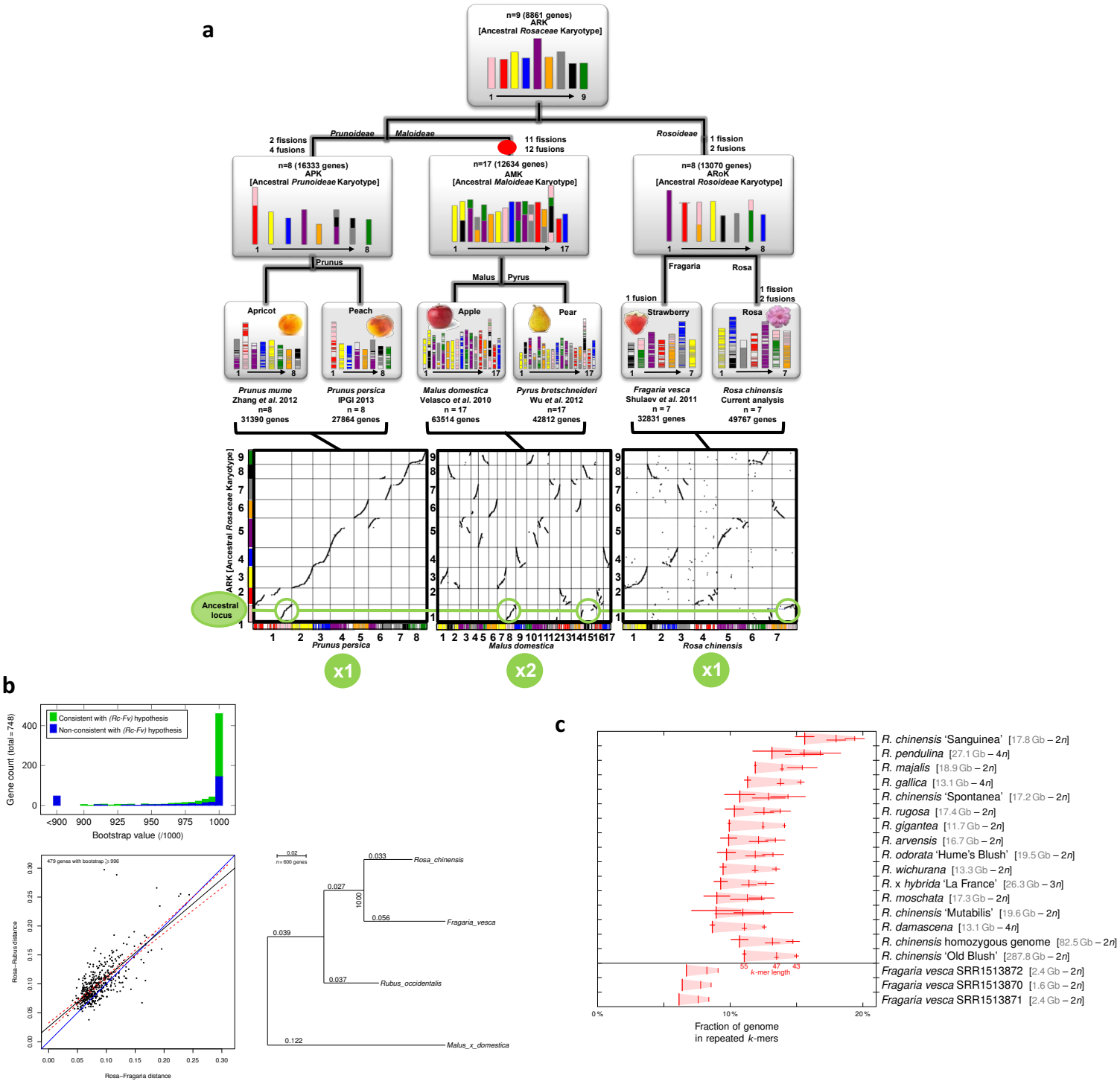
TE families content in homozygous genome

TE families content in heterozygous genome

Supplementary Figure 4. Transposable element annotations.

a, Density of main transposable element superfamilies and genes along *Rosa chinensis* genome assembly. Values are expressed as percentages of sequence length, over 200kb windows. Horizontal brown dashes depict individual caulimoviridae insertions, which cover 1.25% of the genome length.

b, Comparison of transposable element annotations in homozygous (left) and heterozygous (right) genome sequences.

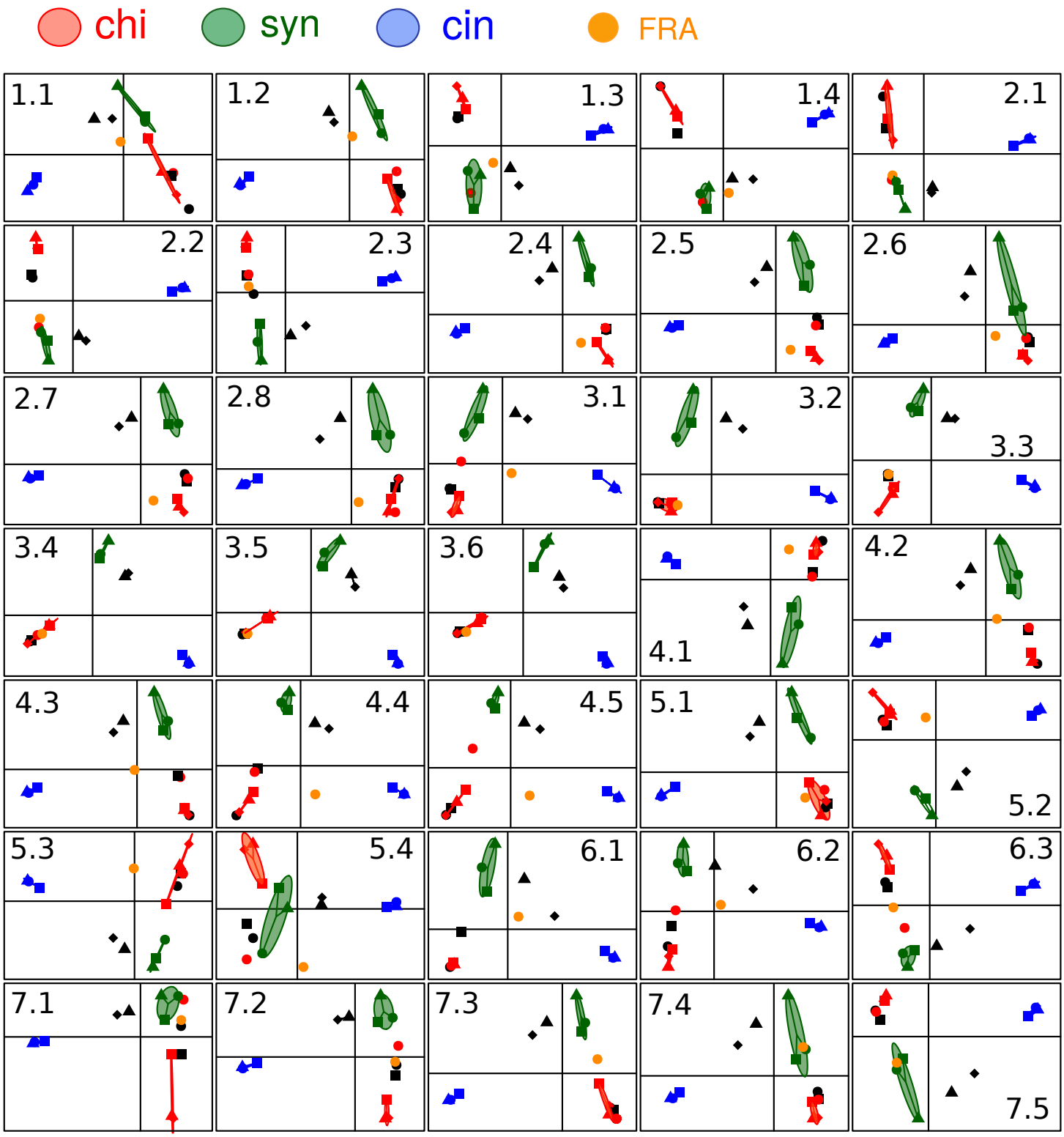


Supplementary Figure 5. Rosaceae evolutionary history.

a, Top: Evolutionary scenario of the modern *Rosaceae* (apricot, peach, apple, pear strawberry, rose) from the ancestral *Rosaceae* karyotype (ARK), ancestral *Prunoideae* karyotype (APK), ancestral *Maloideae* karyotype (AMK) and ancestral *Rosoideae* karyotype (ARoK). The modern genomes are illustrated at the bottom with different colors reflecting the origin from the nine ancestral chromosomes from ARK. Duplication events are shown with red dots on the tree branches, along with the shuffling events (fusions and fissions). Bottom: Complete dot-plot based deconvolution into nine reconstructed CARs (dot-plot y-axis in nine colors) of the observed synteny and paralogy (dot-plot diagonals) between ARK (dot-plot y-axis) and the investigated species (peach, apple and rose as dot-plot x-axis). The complete overview of paralogous and orthologous gene relationships between the modern *Rosaceae* genomes as well as the reconstructed ARK are illustrated in green circles, as case example for ARK protochromosome 1 (pink), for applied translational research.

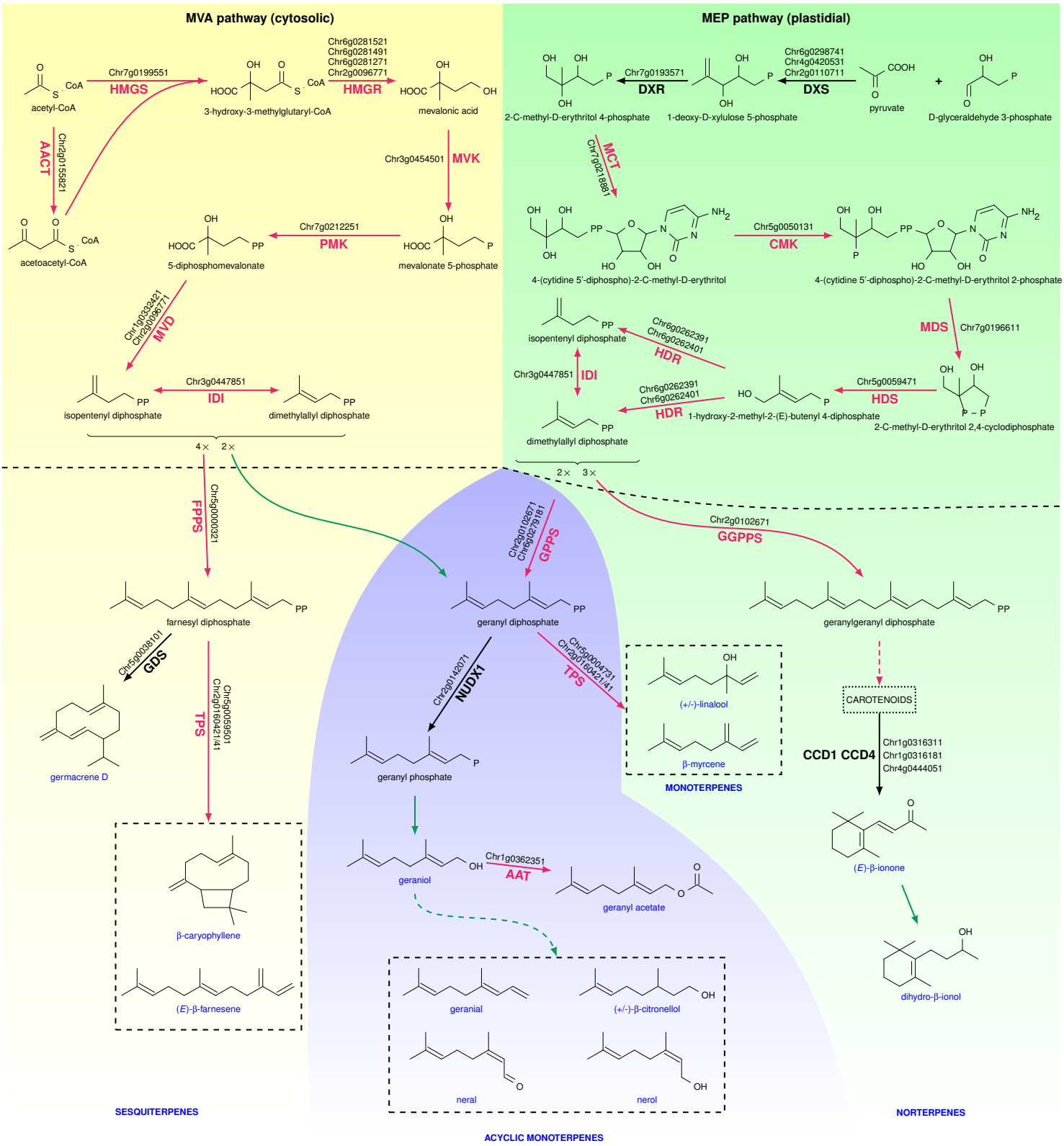
b, Rosoideae radiation: Phylogenetic trees were computed based on the coding sequence of 748 genes from *Rosa chinensis* 'Old Blush', *Rubus occidentalis*, *Fragaria vesca* and, as an outgroup, *Malus* x *domestica*. The base hypothesis was that *Rosa* and *Fragaria* diverged more recently from one another than from *Rubus*. The barplot (top left) shows that most of the trees with high bootstrap values supports this hypothesis, and so does the consensus tree obtained from the concatenation of 600 genes (bottom right), but when considering the *Rosa-Fragaria* and *Rosa-Rubus* distances gene by gene (dot plot in the lower part), we observe that the dots follow the diagonal (in blue) and that the slope is only marginally different from 1 (5% confidence interval in red). This result favor the hypothesis that the three genera diverged approximately at the same time.

c, Comparative *k*-mer analysis between *Rosa* species and *Fragaria vesca* genomes. The fraction of genome represented by repeated *k*-mers of length 55, 47 and 43bp is depicted by vertical bars. *Rosa* datasets were randomly subsampled to 2.4Gb to be comparable to *Fragaria* ones, and the horizontal bars depict the standard deviation over 10 randomizations. The total size of the dataset and ploidy level is given between square brackets for each genotype



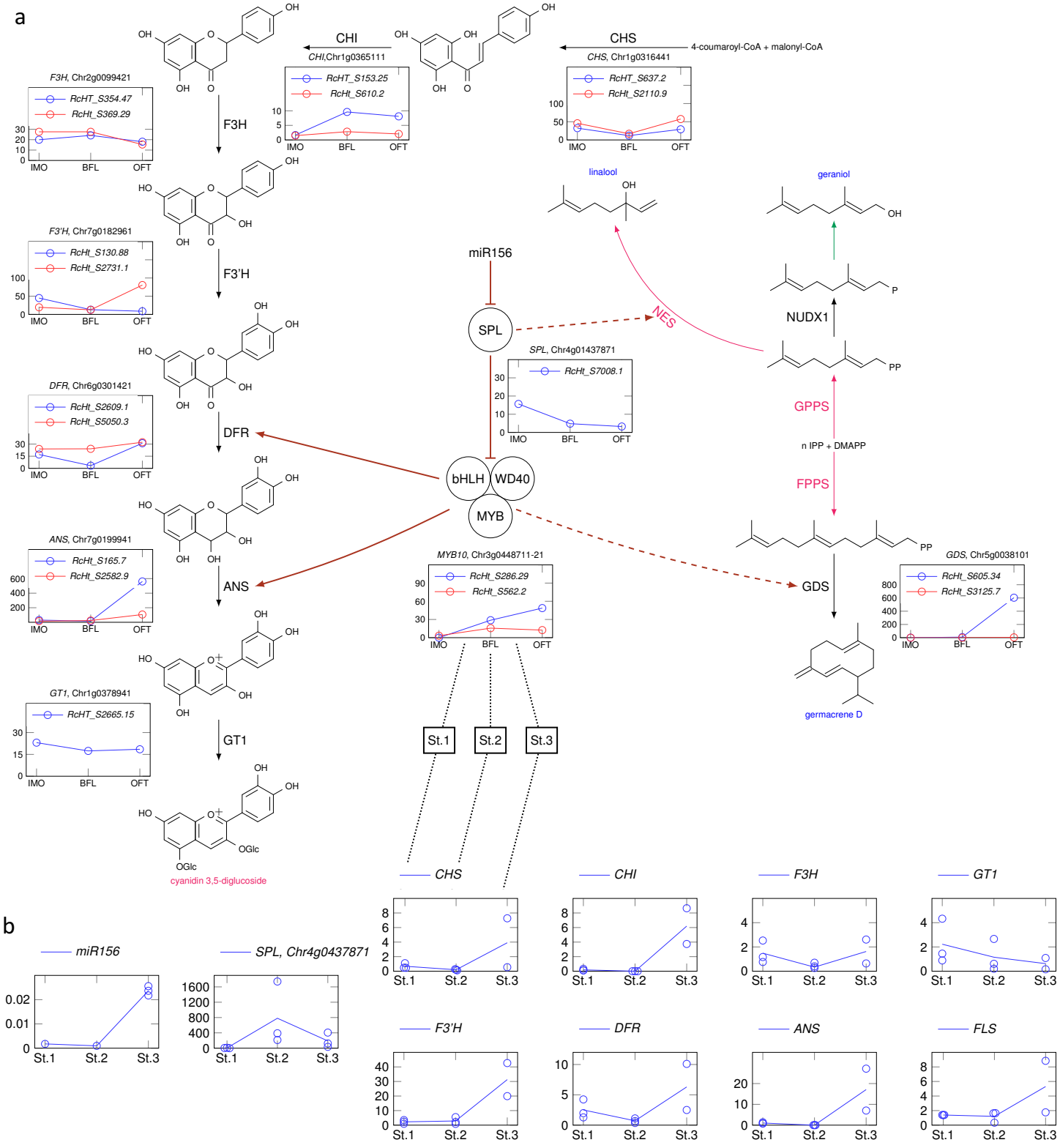
Supplementary Figure 6. Origin of the cultivar *R. x hybrida* 'La France'.

Principal component analyses (PCA) were carried out on genic variants in a dataset of 15 resequenced *Rosa* genotypes. The genome was partitioned into 35 chromosomal segments based on changes in structuration of variants density in the rose cultivars (cf. Figure 2 and Supplementary Data 2). PCA for each segment are represented in the same order as in Figure 2. The Chinenses, Synstylae and Cinnamomeae sections are highlighted with red, green and blue ellipses respectively. The cultivar 'La France' is drawn in orange with other cultivars drawn in black. The X and Y axes represent the first and second component of the PCA and explained 29.29 to 40.53% and 12.07 to 19.89% of the variance, respectively (cf. Supplementary Data 2). The plot was carried out with the s.class function of the R package adegenet. Representation of the different genotypes: *R. gigantea*, red square; *R. chinensis* 'Hume's Blush', red circle; *R. chinensis* 'Sanguinea', red triangle; *R. chinensis* 'Spontanea', red diamond; *R. chinensis* 'Old Blush' heterozygote genotype , black square; *R. chinensis* 'Mutabilis', black circle; *R. gallica*, black triangle; *R. damascena*, black diamond; *R. moschata*, green square; *R. wichurana*, green circle; *R. arvensis*, green triangle; *R. x hybrida* 'La France'; orange circle; *R. pendulina*, blue square; *R. rugosa*, blue circle; *R. majalis*, blue triangle.



Supplementary Figure 7. Terpene biosynthesis pathway in the rose.

Terpenes produced in rose genotypes used in this study, are included. The name of the enzymes acting at different steps and the putative corresponding genes are indicated. Black arrows indicate that the biosynthetic step has already been identified in rose. Red arrows indicate that the biosynthetic step has already been studied in other species, but not in the rose. Green arrows indicate putative steps with unknown enzymes. Dashed arrows indicate several enzymatic steps. Volatile compounds are indicated in blue letters. Molecules in red boxes could be either synthesized by the same enzyme or by different enzymes. The dashed boxes correspond to specific pathways in the plastids (green background) or in the cytosol (yellow background) cellular compartment. The cellular compartment of the pathways in blue background are highly discussed in the bibliography. For example, NUDX1 is cytosolic, but GPP and acyclic monoterpenes biosynthesis are plastidial in other species. AACT, acetoacetyl-CoA thiolase; AAT, alcohol acyl transferase; CCD, carotenoid cleavage dioxygenase; CMK, 4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol kinase; DXR, 1-deoxy-D-xylulose 5-phosphate reductoisomerase; DXS, 1-deoxy-D-xylulose 5-phosphate synthase; FPPS, farnesyl diphosphate synthase; GerD, germacrene D synthase; GPPS, geranyl diphosphate synthase; GGPPS; geranylgeranyl diphosphate synthase; HDR, 1-hydroxy-2-methyl-2-butenyl 4-diphosphate reductase; HDS, 1-hydroxy-2-methyl-2-butenyl 4-diphosphate synthase; HMGR, 3-hydroxy-3-methylglutaryl CoA reductase; HMGS, 3-hydroxy-3-methylglutaryl CoA synthase; IDI, isopentenyl diphosphate isomerase; MCT, 2-C-methyl-D-erythritol 4-phosphate cytidyltransferase; MDS, 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase; MVD, 5-diphosphomevalonate decarboxylase; MEP, 2-C-methyl-D-erythritol 4-phosphate; MVA, mevalonic acid; MVK, mevalonate kinase; NUDX1, nudix hydrolase1; PMK, 5-phosphomevalonate kinase; TPS, terpene synthase.

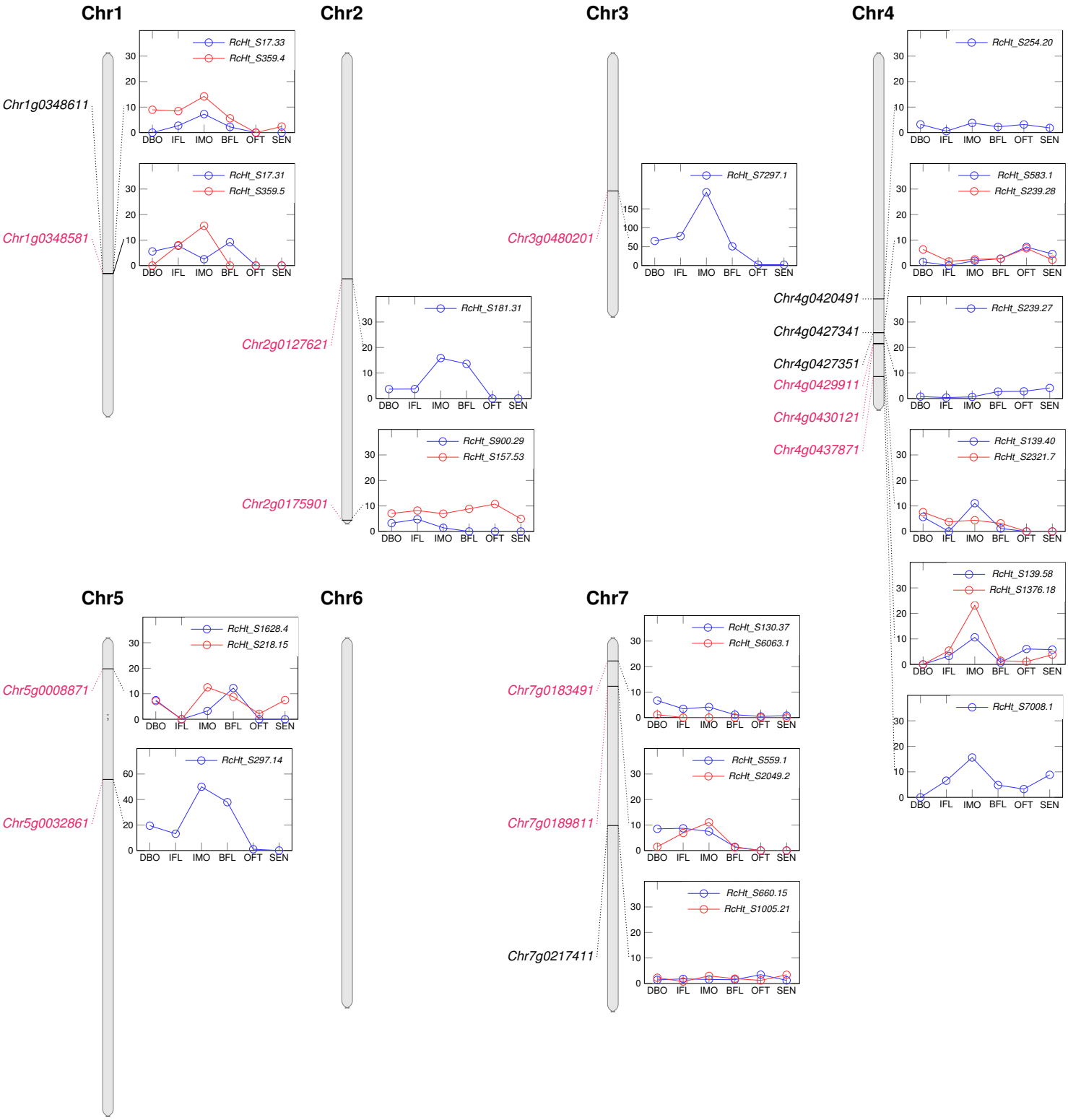


Supplementary Figure 8. Integration of genes in the phenolic pigment and volatile terpenes synthesis pathways.

a, Upper panel : Gene expression at three different floral development stages is shown. IMO=floral meristem and early floral organs, BFL=closed flower, OFT=open flower. Whenever necessary FPKM values are given for each allelic copy of the genes and appear in blue or red. Alleles are identified by their names in the heterozygous annotation. Correspondance between heterozygous and homozygous annotations is given. No expression of *NES* was detected in all analysed tissue.

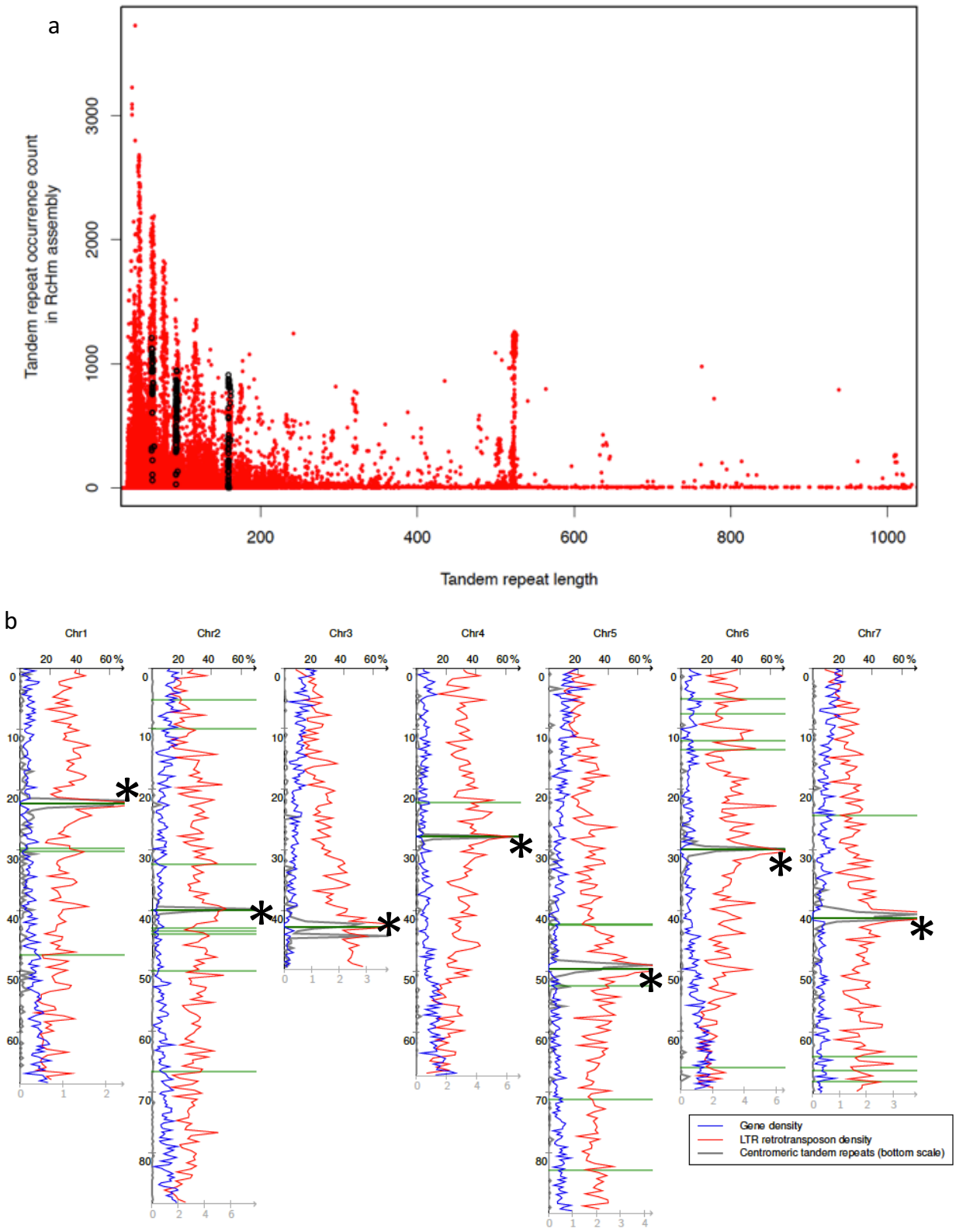
b, Lower panel : RT-qPCR quantification of anthocyanin biosynthesis genes during petal growth and pigmentation.

CHS : CHALCONE SYNTHASE ; CHI : CHALCONE ISOMERASE ; F3H : FLAVANONE 3-HYDROXYLASE ; F3'H : FLAVONOID 3'-HYDROXYLASE ; DFR : DIHYDROFLAVONOL 4-REDUCTASE ; FLS : FLAVONOL SYNTHASE ; ANS : ANTHOCYANIDIN SYNTHASE ; GT1 : ANTHOCYANIDIN 5,3-O-GLUCOSYLTRANSFERASE ; SPL : SQUAMOSA PROMOTER BINDING PROTEIN-LIKE ; GDS : GERMACRENE D SYNTHASE ; GPPS : GERANYL DIPHOSPHATE SYNTHASE ; FPPS : FARNESYL DIPHOSPHATE SYNTHASE ; NES : NEROLIDOL SYNTHASE ; CCD1/4 : CAROTENOID CLEAVAGE DIOXYGENASE 1/4 ; NUDX1 : NUDX HYDROLASE 1



Supplementary Figure 9. ***In silico*** **expression of predicted SPL genes during the course of 'Old Blush' floral development.**

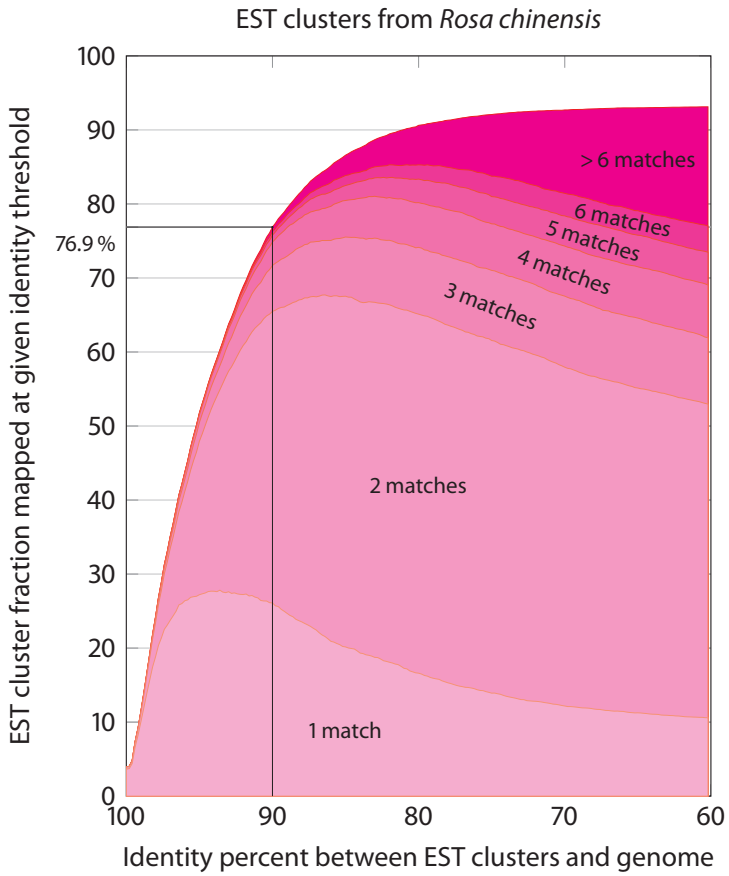
Predicted *SPL,* that are putative targets of miR156, are highlighted in red. These SPL genes are expressed at early floral organ initiation development stages, and their expression decreases during flower opening (OFT). DBO=active axillary buds (vegetative meristem), IFL=floral bud at floral meristem transition, IMO=floral meristem and early floral organs, BFL=closed flower, OFT=open flower, SEN=senescent flower) . Whenever necessary FPKM values are given for each allelic copy of the genes and appear in blue or red histograms. Alleles identifiers for are indicated and correspondence between heterozygous and homozygous annotations is shown in Supplementary Data 1.



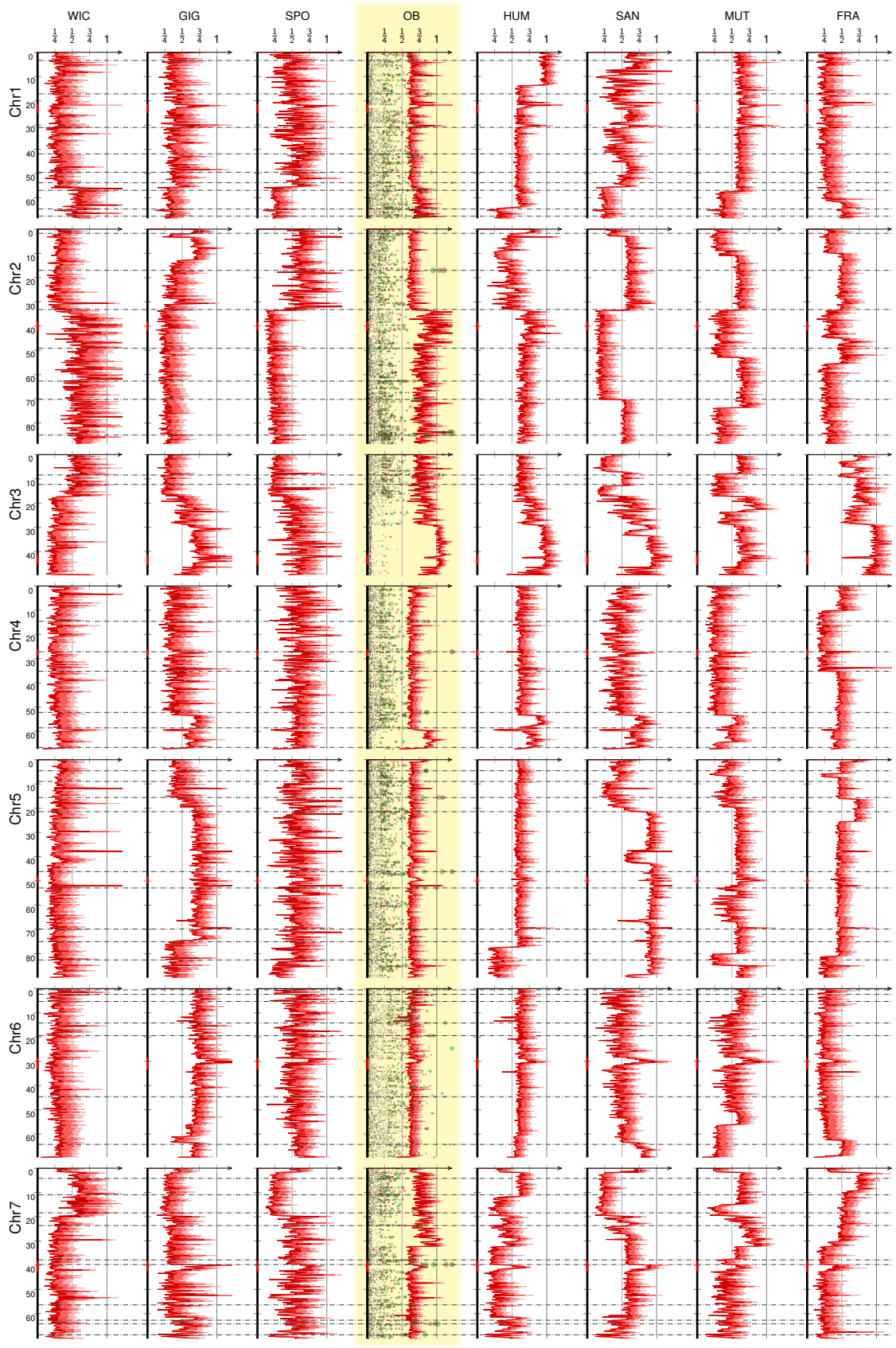
Supplementary Figure 10. **Centromere localization.**

a. Number of occurrences of tandem repeats in the genome, as a function of motif length. Red dots depict all tandem repeats. TRs located in the peaks were considered as candidates for centromeric repeats. Black dots are the final tandem repeats selected as centromere-specific.

b. combined density of centromeric repeats and correlation with gene density and LTR TE density. Green lines show gaps in the assembly. * indicates centromeres position.



Supplementary Figure 11. Mapping of published rose transcripts on *R. chinensis* 'Old Blush' genome sequence. For each identity percent cutoff (horizontal axis), the plot shows the percentage of transcripts having 1 to 6+ matches on *R. chinensis* 'Old Blush' genome sequence (vertical axis). We infer that transcripts having two matches (65.5% of the transcripts at cutoff=90%) correspond to genes for which the two alleles are present in the genome assembly, and transcripts having one match (26.1% at cutoff=90%) correspond to genes for which the two alleles have been assembled as a consensus.



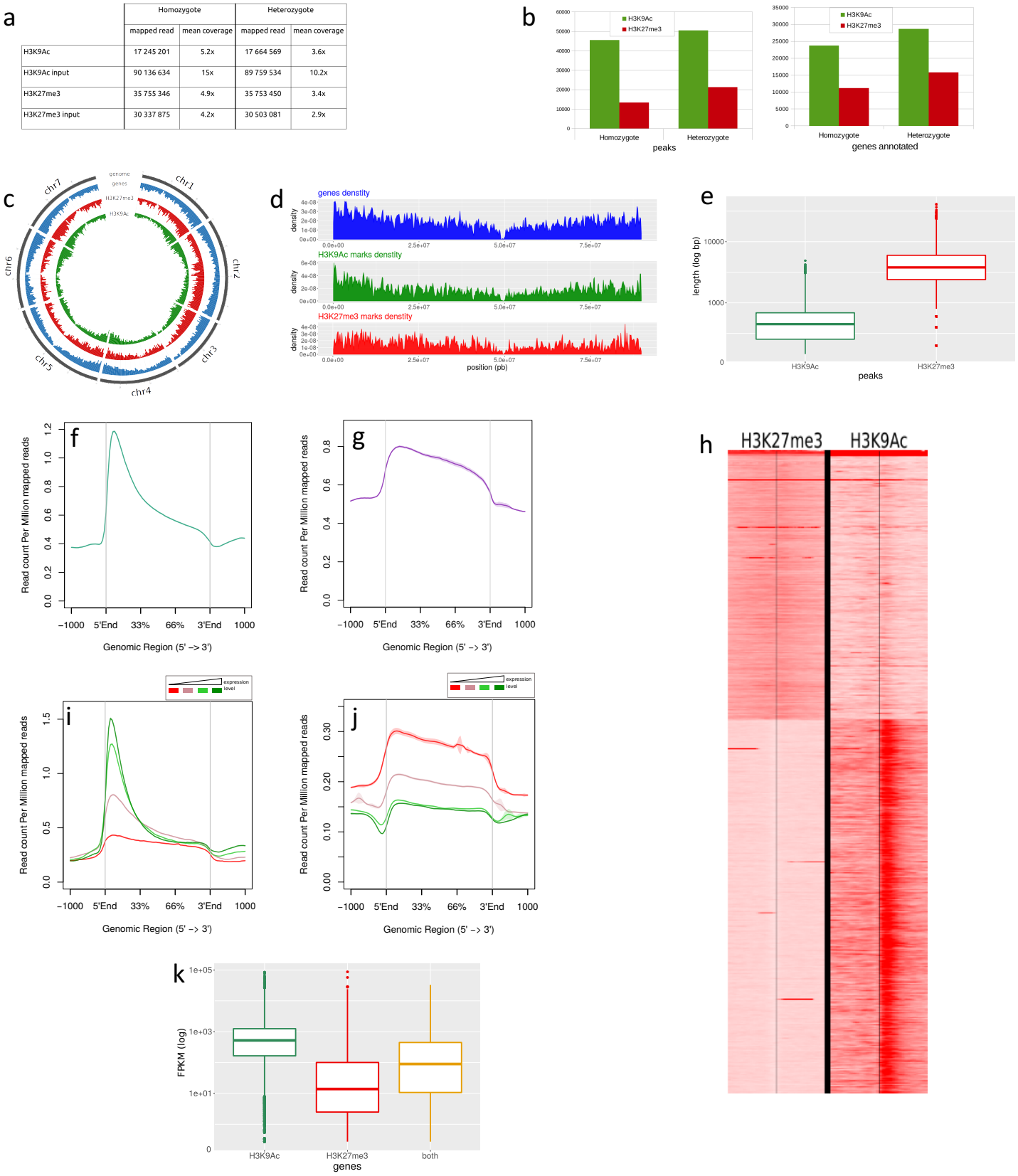
Extended Data Figure 12. Crossing-over localization in *Rc*HzRDP12 genome.

Yellow frame: Crossing-over localization using one-end mapped pairs (OEM). Color dots depict the ratio of OEM pairs over consistent pairs in each 10kb window along the genome. Higher values are on the right. Five Illumina libraries from the heterozygous genome have been used: PE 370bp (green), PE 480bp (brown), PE 630bp (purple), MP 3.3kb (grey), MP 5.4kb (blue). Loci where two or more libraries show a significant enrichment in OEM pairs are considered as candidate crossing-overs and have been depicted with a horizontal dashed line.

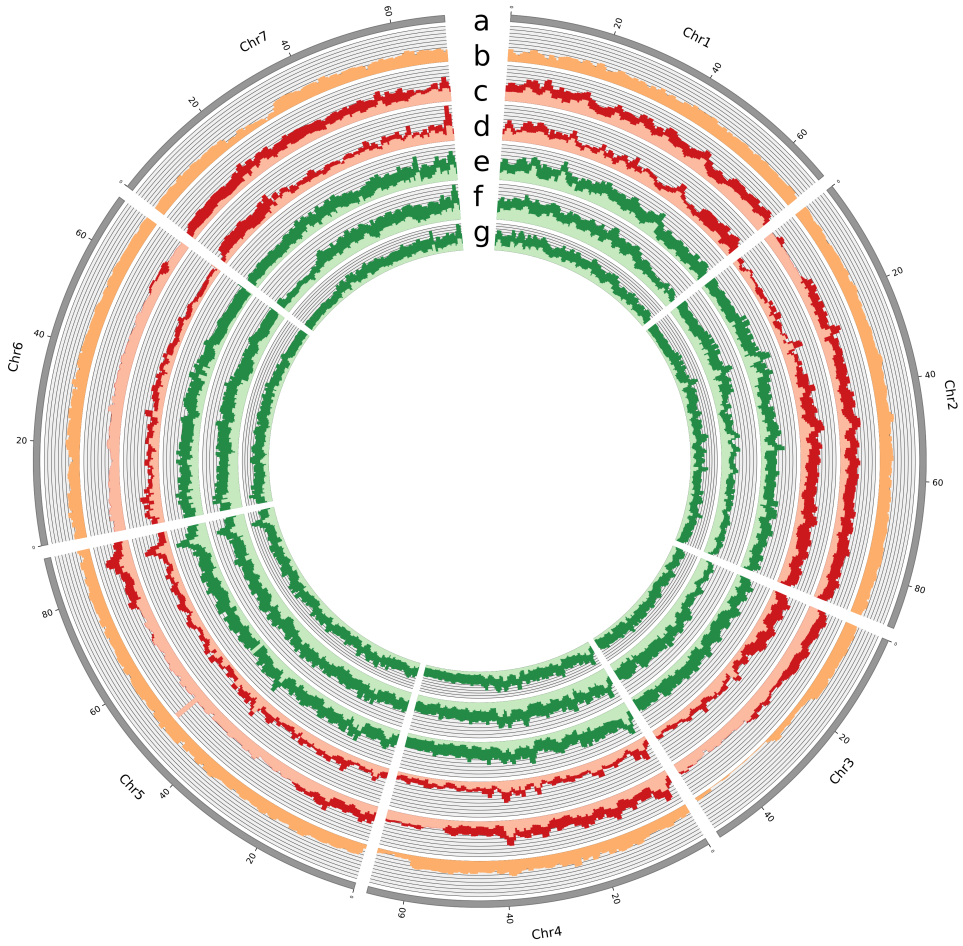
Red plots: Segmental structure of sequence conservation between *Rosa* species. Red curves along the chromosomes depict the level of sequence conservation between the homozygous genome and 8 *Rosa* genotypes, including 'Old Blush' (Supplementary Notes 8). A conservation value of 1 means that the sequences are completely identical to the homozygous one, in both haplotypes of the resequenced genotypes. Conservation can be higher than 1 at a low stringency due to repeated sequences. Centromeres are displayed as red lines on the chromosomes.

a

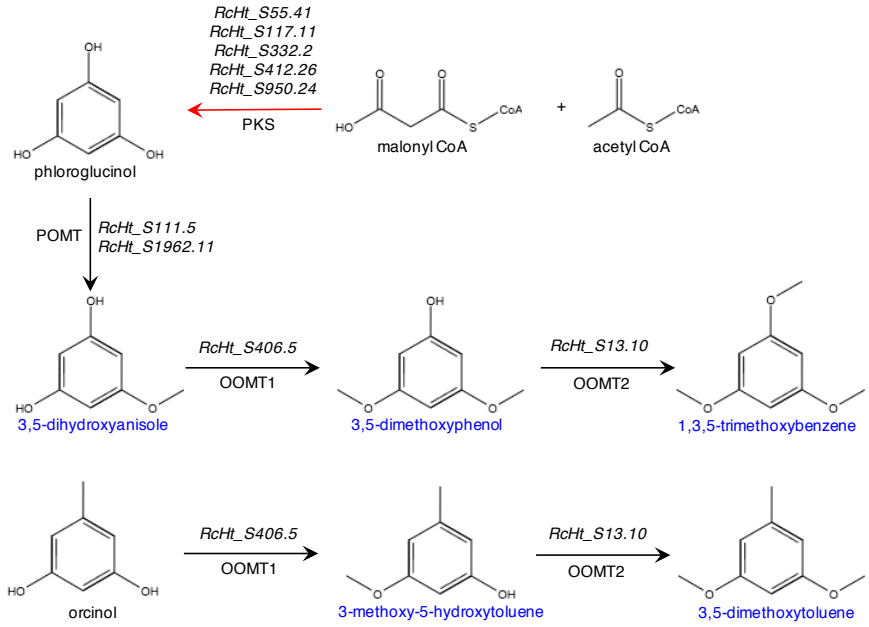
	Homozygote		Heterozygote	
	mapped read	mean coverage	mapped read	mean coverage
H3K9Ac	17 245 201	5.2x	17 664 569	3.6x
H3K9Ac input	90 136 634	15x	89 759 534	10.2x
H3K27me3	35 755 346	4.9x	35 753 450	3.4x
H3K27me3 input	30 337 875	4.2x	30 503 081	2.9x



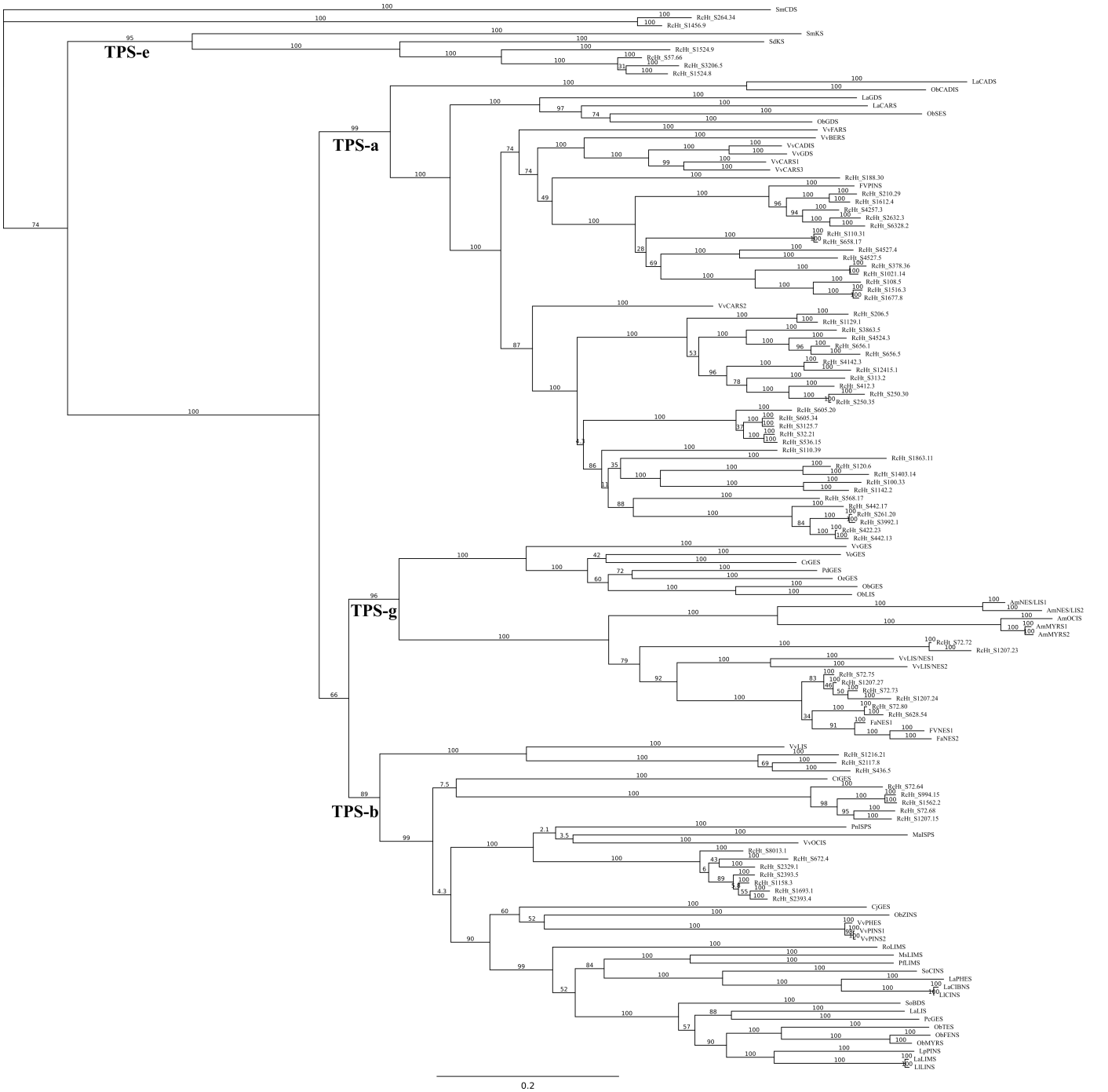
Supplementary Figure 13. a, ChIP-seq mapping metrics **b,** Number of detected peaks for H3K9Ac and H3K27me3 marks (left). Number of annotated genes for H3K9ac and H3K27me3 marks (right). **c,** Distribution of mapped reads for H3K27me3 (red shades) and H3K9ac (green shades) along the 7 rose chromosomes. Local peak densities of each epigenetic mark were plotted against the genetic distance (gray) and annotation of transcripts (blue). **d,** H3K27me3 and H3K9ac distribution at the chromosome level. Distribution of annotated genes (blue, upper panel), H3K9ac marks (green, medium panel) and H3K27me3 marks (red, bottom panel) in flowers are plotted along the chromosome 5. **e,** Box plot of H3K9ac peaks length (green) and H3K27me3 peaks length (red). **f, g,** Average tag density profile of H3K27me3 and H3K9ac along the gene body. ChIP-Seq densities of equal bins were plotted along the gene body and 2-kb region flanking the TSS or the TES. **h,** Heat map representing the tag density distribution of H3K27me3 and H3K9ac across all genes and a 2kb flank. **i, j,** Correlation of H3K27me3, H3K9ac and gene expression level. All the rose protein-coding genes were divided in 4 quantiles according to their gene expression levels (lowest and highest expression level corresponding to red and green, respectively). For each quantile the number of H3K27me3 and H3K9ac mapped reads was averaged and plotted along the gene body and 1-kb region flanking the TSS or the TES. **k,** Boxplot showing the mean expression value of genes marked by H3K9ac, H3K27me3 or both H3K9ac and H3K27me3



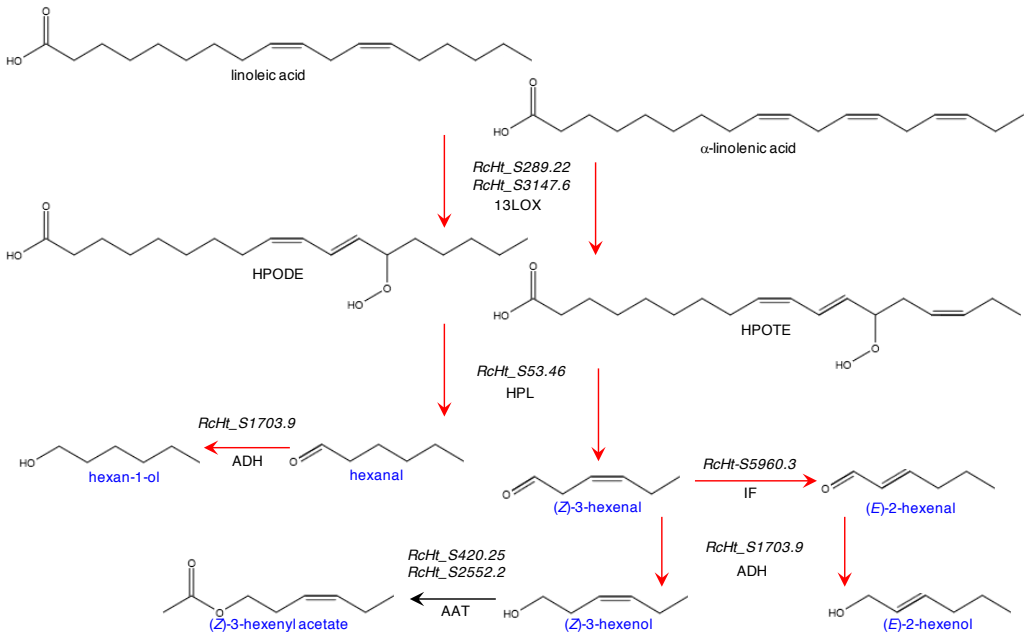
Supplementary Figure 14. Density of genic variants in 1 Mb sliding windows in resequenced genotypes.
a, Schematic representation of the pseudomolecules of the double haploid reference genome.
b, *R. chinensis* 'Old Blush' (heterozygote genotype), in orange. **c,** *R. gigantea*. **d,** *R. chinensis* 'Spontanea'. **e,** *R. moschata*. **f,** *R. wichurana*. **g,** *R. arvensis*. Heterozygote variants are in light shade, homozygote variants are in dark shade. Genotypes of the Chinenses and Synstylae sections are drawn in red and green, respectively.



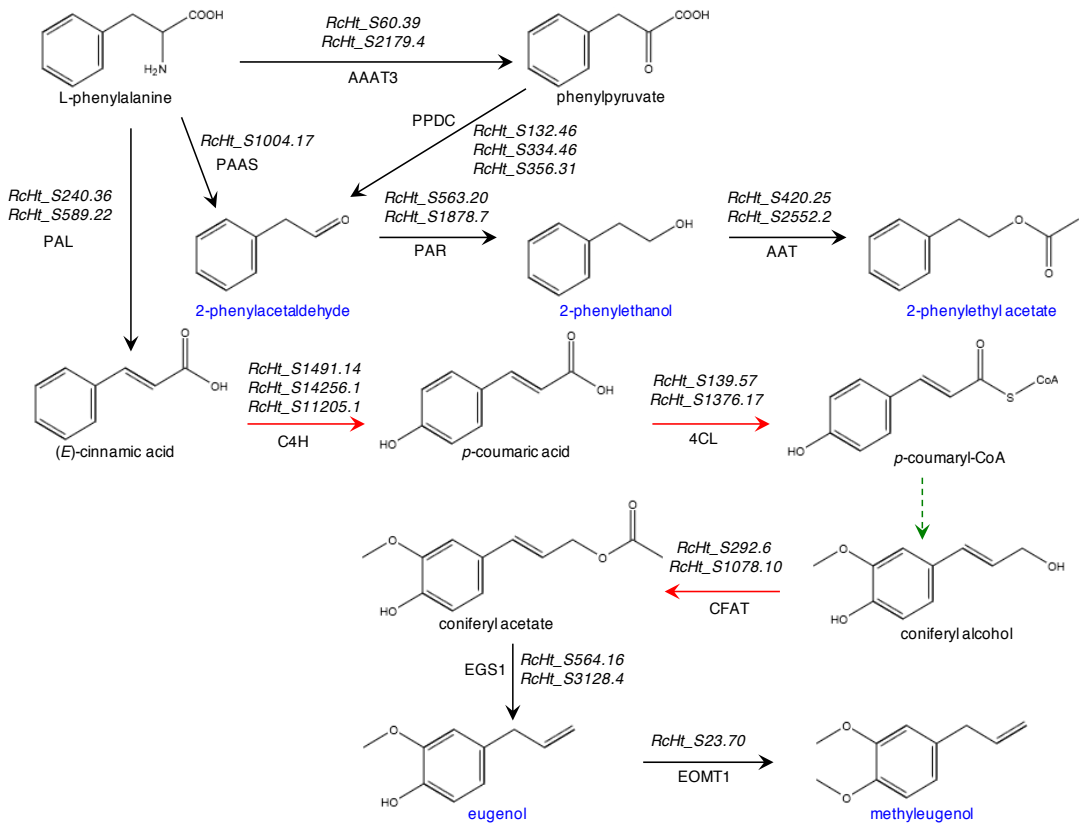
Supplementary Figure 15. Phenolic methyl ether biosynthesis pathway in rose. The name of enzymes acting in different steps and the putative corresponding genes are indicated. Black arrows indicate biosynthetic step that have been identified in rose. Red arrow indicates that the biosynthetic step has been studied in other species, but not in the rose. Volatile compounds are indicated in blue letters. OOMT1 and OOMT2: orcinol *O*-methyl transferase 1 and 2; PKS: polyketide synthase; POMT: phloroglucinol *O*-methyl transferase.



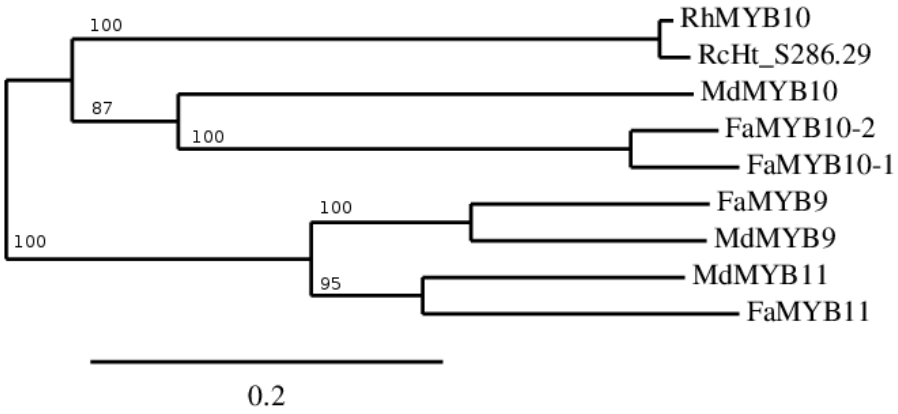
Supplementary Figure 16. Phylogenetic analysis of *R. chinensis* 'Old blush' putative TERPENE SYNTHASES (TPS). Using Geneious software (https://www.geneious.com), amino acid sequences from the heterozygous genome were aligned with Muscle[91] and the tree constructed using the Neighbor-Joining method with 1000 iterations. The bootstrap values >50% are shown; the scale bar (0.2) corresponds to the number of amino acid substitutions per site. TPS from other species, whose function has been demonstrated *in vitro,* were included in the analysis. Am, *Antirrhinum majus*; Cj, *Citrus jambhiri*; Cr, *Catharanthus roseus*; Ct, *Cinnamomum tenuifolium*; Fa, *Fragaria ananassa*; Fv, *Fragaria vesca*; La, *Lavandula angustifolia*; Ll, *Lavandula latifolia*; Lp, *Lavandula pedunculata*; Ma, *Melaleuca alternifolia*; Ms, *Mentha spicata*; Ob, *Ocimum basilicum*; Oe, *Olea europaea*; Pc, *Perilla citriodora*; Pd, *Phyla dulcis*; Pf, *Perilla frutescens*; Pn, *Populus nigra*; Ro, *Rosmarinus officinalis*; Sm, *Salvia miltiorDOGrhiza*; Sd, *Scoparia dulcis*; Vo, *Valeriana officinalis*; Vv, *Vitis vinifera*. GenBank accession numbers are as follows: AmMYRS1, AAO41726.1; AmMYRS2, AAO41727.1; AmNES/LIS1, ABR24417.1; AmNES/LIS2, ABR24418.1; AmOCIS, AAO42614.1; CjGES, BAM29049.1; CrGES, AFD64744.1; CtGES, CAD29734.2; FaNES1, POCV94.1; FANES2, POCV95.1; FVNES1, POCV96.1; FVPINS, O23945.2; LaCADS, AGL98418.1; LaCARS, AGL98419.1; LaGDS, AGL98420.1; LaLIMS, ABB73044.1; LaLIS, ABB73045.1; LaPHES, ADQ73631.1; LlCINS, AFL03422.1; LlLIS, ABD77417.1; LpPINS, AGN72799.1; MaISPS, AAP40638.1; MsLIMS, AAC37366.1; ObCADIS, AAV63787.1; ObFENS, AAV63790.1; ObGDS, AAV63786.1; ObGES, AAR11765.1; ObLIS, AAV63789.1; ObMYRS, AAV63791.1; OeGES, AFI47926.1; PcGES, ABB30216.1; PdGES, ADK62524.1; PfLIMS, AAG31438.1; PnISPS, ADV58934.1; RoLIMS, ABD77416.1; SdKS, AEF33360; SmCDS, ABV57835.1; SmKS, ABV08817.1; SoBDS, AAC26017.1; SoCINS, AAC26016.1; VoGES, AHE41084.1; VvBERS, ADR74195.2; VvCADIS, ADR74199.1; VvCARS1, ADR74192.1; VvCARS2, ADR74193.1; VvCARS3; ADR74194.1; VvFARS, ADR74198.1; VvGDS, ADR74197.1; VvGES , NP001267920.1; VvLIS, ADR74209.1; VvLIS/NES1, ADR74210.1; VvLIS/NES2, ADR74211.1; VvOCIS, ADR74204.1; VvPHES, ADR74201.1; VvPINS1, ADR74202.1; VvPINS2, ADR74203.1.



Supplementary Figure 17. Green leaf volatile biosynthesis pathway in rose. The name of the enzymes acting at different steps and the putative corresponding genes of the rose genome are indicated. Black arrows indicate biosynthetic steps that have been identified in the rose. Red arrows indicate biosynthetic step that have been reported in plants, but not in rose. Volatile compounds are indicated in blue letters. HPODE, (13*S*)-hydroperoxy-(9*Z*,11*E*)-octadecadienoic acid; HPOTE, (13*S*)-hydroperoxy-(9*Z*,11*E*,15*Z*)-octadecatrienoic acid; ADH: alcohol dehydrogenase; AAT: alcohol acyltransferase; HPL: hydroperoxide lyase; LOX: lipoxygenase.



Supplementary Figure 18. Benzenoids and phenylpropanoids biosynthesis pathway in rose. The name of enzymes acting at different steps and the putative corresponding genes in the rose genome are indicated. Black arrows indicate biosynthetic steps that have been identified in the rose. Red arrows indicate biosynthetic step that have been reported in plants, but not in rose. Green arrows indicate putative steps with an unknown enzyme. Dashed arrows indicate several enzymatic steps. Volatile compounds are indicated in blue letters. Benzaldehyde and benzyl alcohol are not illustrated because enzymes are note known, but they could derive from *t*-cinnamic acid. AAT: alcohol acyltransferase; AAAT3: aromatic amino acid aminotransferase; CFAT: coniferyl alcohol acyltransferase; C4H, cinnamoyl-CoA hydratase-dehydrogenase; 4CL, putative 4-coumarate-CoA ligase; EGS1: eugenol synthase; EOMT: eugenol O-methyltransferase; PAAS: phenylacetaldehyde synthase gene; PAL: phenylalanine ammonia lyase; PAR: phenylacetaldehyde reductase gene; PPDC: phenylpyruvic acid decarboxylase.

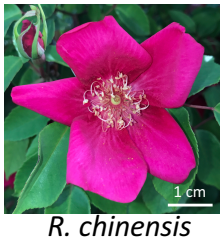


Supplementary Figure 19. Phylogenetic analysis of *R. chinensis* 'Old blush' *RhMYB10*. BioNJ software[1] was used. MYB amino acid sequences from the heterozygous genome were aligned with Muscle and the tree constructed using the Neighbor-Joining method with 1000 iterations. The bootstrap values >50% are shown; the scale bar (0.2) corresponds to the number of amino acid substitutions per site. MYB genes from *Fragaria* (Fa) and *Malus domestica* (Md) known to activate anthocyanin biosynthesis in strawberry and apple were included in the analysis. Protein accession numbers are provided in Supplementary Data 10.faa.

1. Gascuel, O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol* **14**, 685-95 (1997).

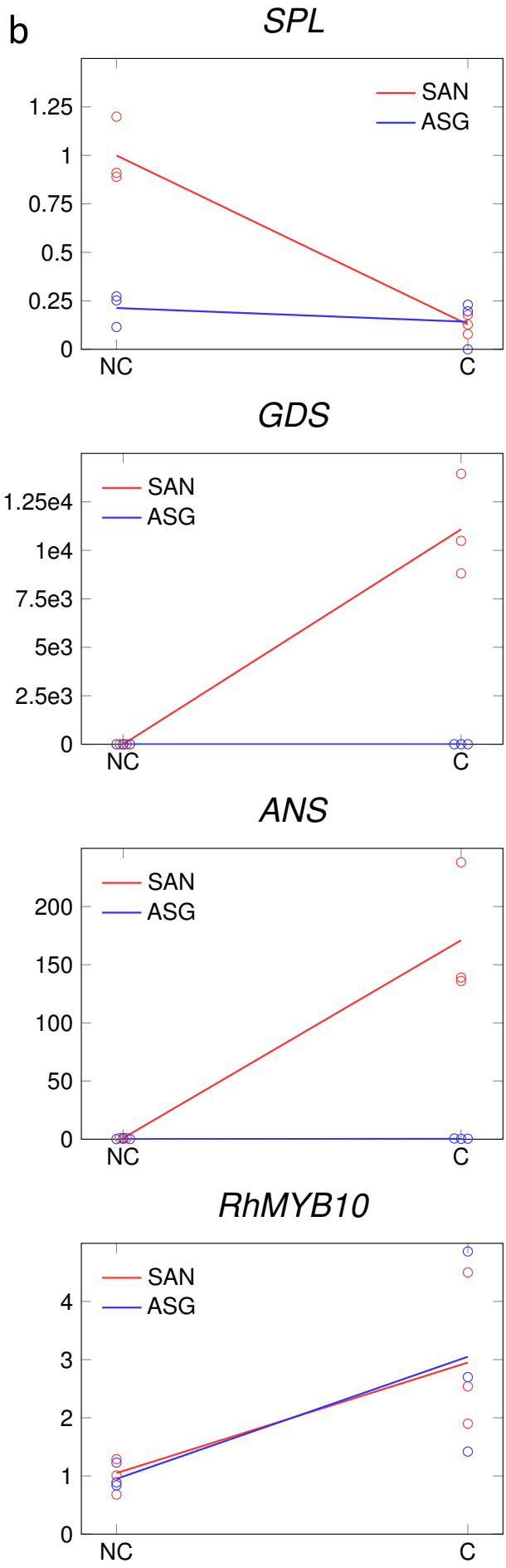
a

R. hybrida
'Alister Stella Gray'



R. chinensis
'Sanguinea'

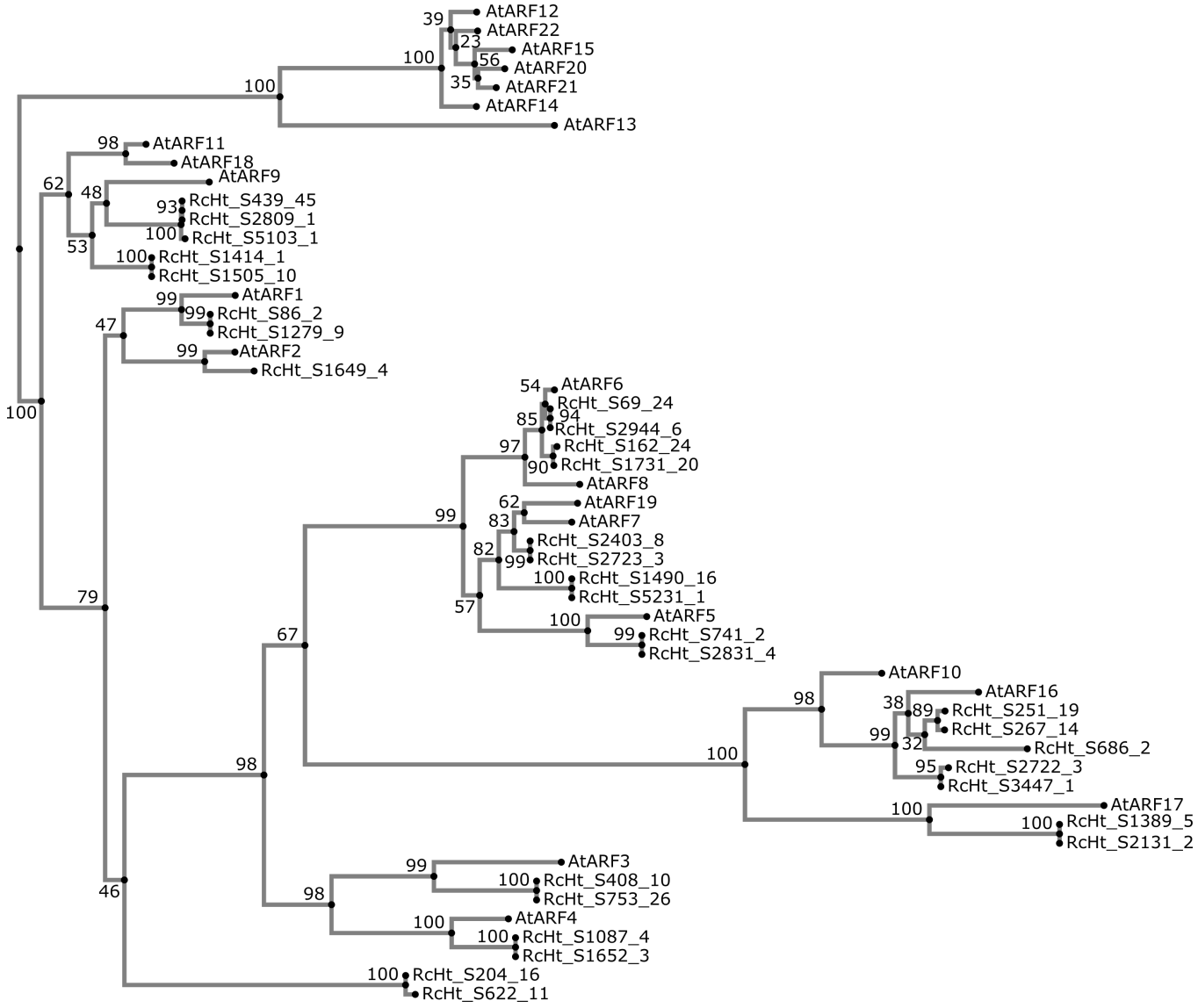
b



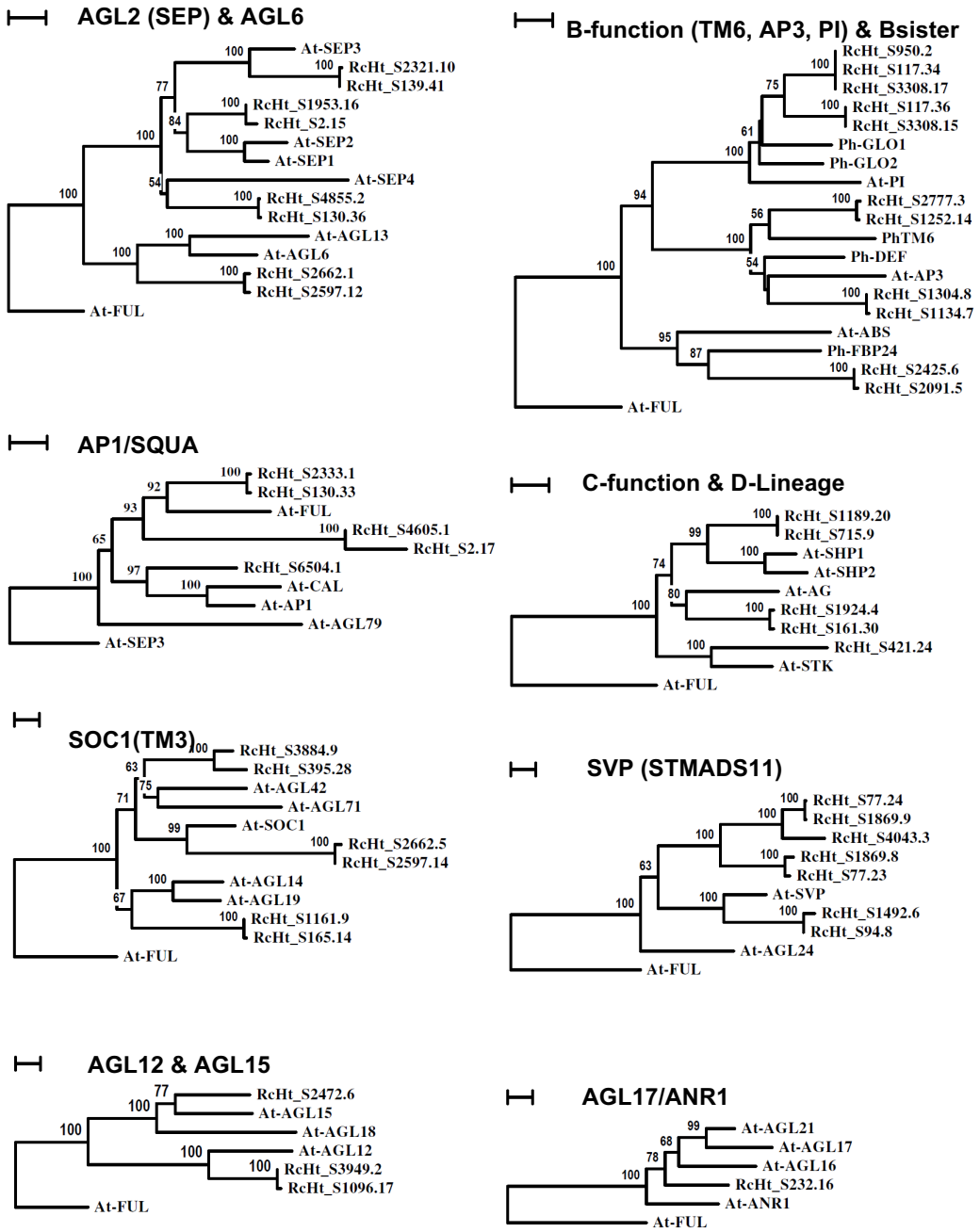
Supplementary Figure 20. Expression of *SPL9*, *ANS*, *GDS* and *MYB10* genes in rose cultivars exhibiting contrasted color.

a, Flowers of *R. hybrida* 'Alister Stella Gray' and *R. chinensis* 'Sanguinea'.

b, RT-qPCR were performed on petals harvested at two successive stages corresponding to non-colored (NC) flower buds and colored (C) opening flowers. SAN-NC and SAN-C: Non-colored and colored flowers, respectively, of *R. chinensis* 'Sanguinea'. ASG-NC and ASG-C: Non-colored and colored flowers, respectively, of *R. hybrida* 'Alister Stella Gray'.



Supplementary Figure 21. Neighbor-joining analysis of *R. chinensis* and *Arabidopsis thaliana* Auxin Response Factor gene family. Local bootstrap probabilities are indicated for branches with >50% support, based on 1000 replicates. Sequence prefixes: RcHt*: R. chinensis*; At: *Arabidopsis thaliana.* Distance scale bars correspond to 0.1 substitution /site.



Supplementary Figure 22. Neighbor-joining analysis of *Rosa chinensis* and *Arabidopsis thaliana* type II MADS-box proteins. Local bootstrap probabilities are indicated for branches with >50% support, based on 1000 replicates. Sequence prefixes: RcHt*: R. chinensis*; At: *Arabidopsis thaliana*; Ph: *Petunia hybrida.* Distance scale bars correspond to 0.1 substitution/site.

Premeiotic pathway

S6 Kinase	**KPK1_ARATH (At3g08730)**	**KPK2_ARATH (At3g08720)**
RchiOBHmChr3g0460281	(65.39%; 471; 0.0)	(67.57%; 478; 0.0)
RchiOBHmChr7g0180021	(55.11%; 450; 5e-169)	(58.05%; 441; 0.0)
RBR	**RBR1_ARATH (At3g12280)**	
RchiOBHmChr1g0326641	(70.65%; 1029; 0.0)	
E2F	**E2FA_ARATH (At2g36010)**	**E2FB_ARATH (At5g22220)**
RchiOBHmChr3g0475421	(47.88%; 518; 7e-136)	(51.07%; 466; 1e-129)
RchiOBHmChr5g0047281	(48.31%; 394; 4e-100)	(57.14%; 416; 4e-143)
40S Ribosomal protein S	**RS61_ARATH (At4g31700)**	**RS62_ARATH (At5g10360)**
RchiOBHmChr2g0112471	(88.76%; 249; 5e-153)	(89.16%; 249; 6e-160)
RchiOBHmChr7g0242901	(91.2%; 250; 2e-166)	(90.76%; 249; 4e-165)
CDK B1	**CKB11_ARATH (At3g54180)**	**CKB12_ARATH (At2g38620)**
RchiOBHmChr3g0459391	(84.47%; 309; 0.0)	(83.28%; 311; 0.0)
RchiOBHmChr5g0011221	(42.17%; 313; 6e-72)	(41.59%; 315; 1e-73)
RchiOBHmChr6g0279771	(42.81%; 327; 2e-75)	(42.86%; 329; 5e-77)
GSL8	**CALSA_ARATH (At2g36850)**	**CALS9_ARATH (At3g07160)**
RchiOBHmChr7g0191411	(75.30%; 1968; 0.0)	(59.67%; 1976; 0.0)
RchiOBHmChr5g0080681	(61.50%; 1922; 0.0)	(76.20%; 1912; 0.0)
SMT 2-3	**SMT2_ARATH (At1g20330)**	**SMT3B_ARATH (At1g76090)**
RchiOBHmChr6g0303691	(86.20%; 355; 0.0)	(82.54%; 355; 0.0)
RchiOBHmChr4g0386051	(84.23%; 355; 0.0)	(80.85%; 355; 0.0)

Switch mitosis-meiosis

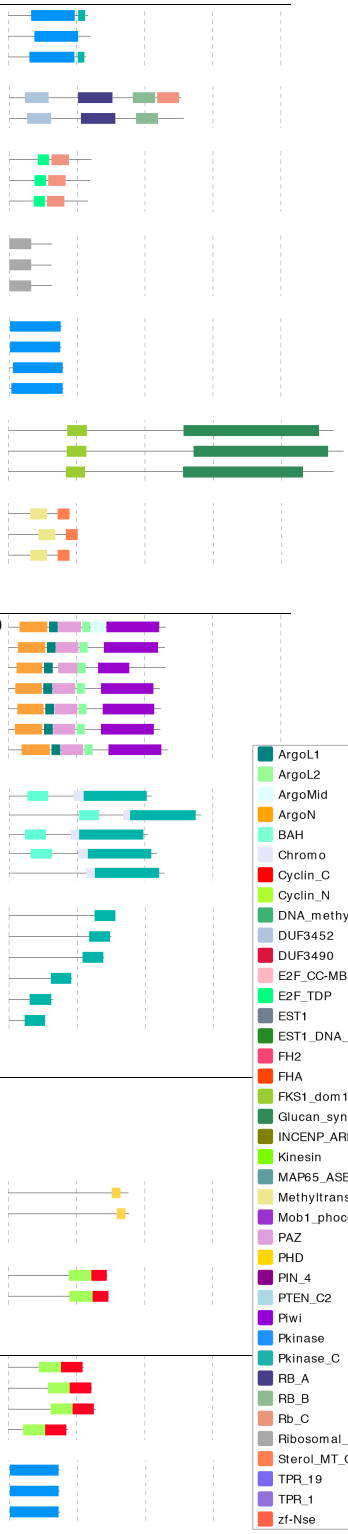
AGO4 clade	**AGO4_ARATH (At2g27040)**	**AGO6_ARATH (At2g32940)**	**AGO9_ARATH (At5g21150)**	**AGO8_ARATH (At5g21030)**
RchiOBHmChr1g0330301	(61.92%; 906; 0.0)	(53.56%; 900; 0.0)	(61.26%; 926; 0.0)	(53.92%; 887; 0.0)
RchiOBHmChr2g0157951	(52.25%; 946; 0.0)	(57.5%; 919; 0.0)	(53.07%; 917; 0.0)	(49.83%; 886; 0.0)
RchiOBHmChr2g0157971	(58.82%; 867; 0.0)	(63.05%; 866; 0.0)	(60.65%; 864; 0.0)	(54.46%; 869; 0.0)
RchiOBHmChr2g0157981	(60.69%; 870; 0.0)	(63.52%; 880; 0.0)	(60.14%; 883; 0.0)	(54.33%; 854; 0.0)
RchiOBHmChr4g0412801	(63.77%; 875; 0.0)	(55.18%; 869; 0.0)	(62.43%; 872; 0.0)	(56.23%; 873; 0.0)
RchiOBHmChr5g0049901	(74.52%; 882; 0.0)	(56.34%; 872; 0.0)	(69.52%; 896; 0.0)	(63.66%; 885; 0.0)
DNA methyltransferase	**CMT3_ARATH (At1g69770)**	**CMT1_ARATH (At1g80740)**	**CMT2_ARATH (At4g19020)**	
RchiOBHmChr7g0212911	(48.02%; 783; 0.0)	(45.33%; 781; 0.0)	(60.25%; 815; 0.0)	
RchiOBHmChr4g0398431	(52.56%; 841; 0.0)	(48.58%; 776; 0.0)	(49.62%; 800; 0.0)	
RchiOBHmChr4g0444071	(51.57%; 828; 0.0)	(45.15%; 824; 0.0)	(48.49%; 796; 0.0)	
RchiOBHmChr2g0085701	(53.47%; 821; 0.0)	(48.54%; 787; 0.0)	(48.76%; 804; 0.0)	
	DRM2_ARATH (At5g14620)	**DRM1L_ARATH (At5g15380)**		
RchiOBHmChr2g0172381	(52.98%; 540; 0.0)	(53.12%; 537; 0.0)		
RchiOBHmChr2g0136781	(51.91%; 549; 0.0)	(47.31%; 613; 0.0)		
RchiOBHmChr1g0322691	(55.16%; 368; 2e-134)	(55.26%; 380; 1e-134)		
RchiOBHmChr1g0323101	(58.26%; 242; 7e-89)	(63.79%; 232; 7e-99)		
RchiOBHmChr2g0136801	(65.84%; 202; 5e-91)	(64.90%; 208; 1e-90)		

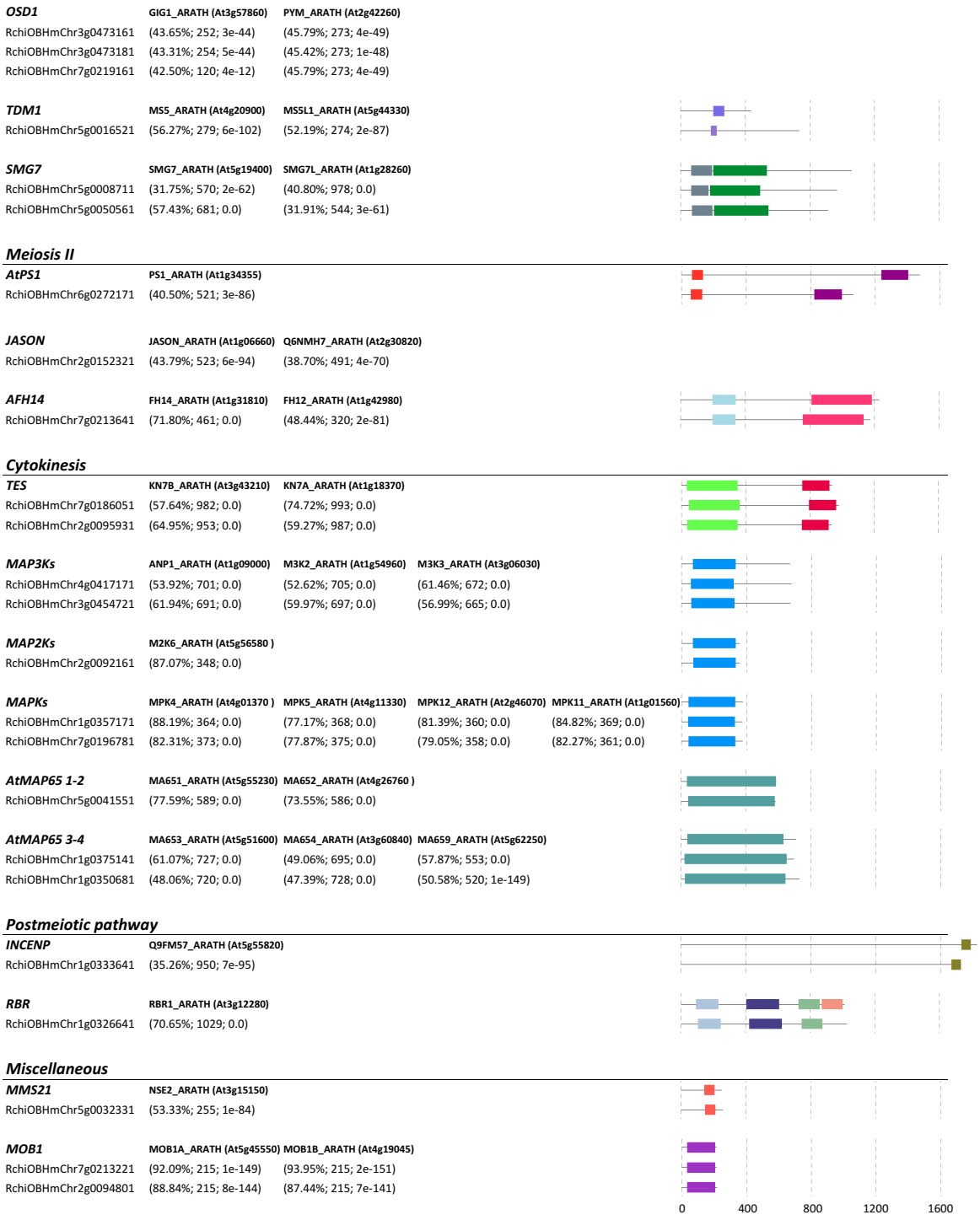
Meiosis I

DYAD/SWI1	**DYAD_ARATH (At5g51330)**	**F4KE84_ARATH (At5g23610)**		
RchiOBHmChr1g0373991	(37.74%; 575; 4e-112)	(37.44%; 211; 3e-37)		
RchiOBHmChr2g0120161	(35.60%; 663; 3e-103)	(41.18%; 204; 2e-36)		
RchiOBHmChr3g0484951	(39.90%; 203; 1e-34)	(37.07%; 375; 5e-63)		
DUET/MMD1	**MMD1_ARATH (At1g66170)**	**Y2181_ARATH (At2g01810)**	**F4IMC6_ARATH (At2g07714)**	
RchiOBHmChr3g0492331	(54.24%; 695; 0.0)	(41.32%; 726; 0.0)	(54.12%; 170; 8e-49)	
RchiOBHmChr3g0450161	(45.60%; 511; 1e-132)	(36.83%; 505; 6e-100)	(55.56%; 126; 2e-41)	
SDS	**CCSDS_ARATH (At1g14750)**			
RchiOBHmChr4g0442421	(45.21%; 574; 1e-125)			

Meiosis I-II transition

TAM/CYCA1;2	**CCA12_ARATH (At1g77390)**	**CCA11_ARATH (At1g44110)**	
RchiOBHmChr6g0296431	(50.11%; 474; 6e-150)	(59.76%; 497; 0.0)	
RchiOBHmChr6g0296421	(46.23%; 491; 1e-138)	(55.81%; 430; 2e-167)	
RchiOBHmChr2g0095961	(60.78%; 288; 2e-121)	(63.12%; 301; 8e-140)	
CDKA;1	**CDKA1_ARATH (At3g48750)**		
RchiOBHmChr1g0333011	(84.35%; 294; 0.0)		
RchiOBHmChr3g0489001	(86.05%; 294; 0.0)		





Supplementary Figure 23. Genetic pathways involved in diploid gamete formation. Putative Rose orthologues of *Arabidopsis* genes involved in diploid gamete formation were searched for by reciprocal best-hits blast approach. Both genes and their co-orthologues (MetaPhOrs database) were taken into account. Identity percentage, alignment length and E-value are indicated by brackets. For both Rose and main *Arabidopsis* proteins, domain structure is displayed from Pfam 31.0 database HMM search results.

Supplementary Table 1. Summary of transposable element and repeat annotation.

TE families*	Homozygous genome coverage (%)	Heterozygous genome coverage (%)
Class I – RNA retrotransposons		
RLC-Copia	9.968	8.749
RLG-Gypsy	12.906	9.831
RIX-LINE	6.793	8.182
Potential-RSX-SINE	0.153	0.191
RXX-ClassI	0.012	0.013
RXX-LARD	1.692	0.700
RXX-TRIM	0.091	0.105
Class II – DNA transposons		
DTX-TIR	9.235	8.918
DXX-MITE	1.342	1.374
DXX-other ClassII	0.692	0.695
DHX-Helitron	0.400	0.369
Chimeric	7.513	4.463
Unclassified	7.868	7.335
Caulimoviridae	1.247	0.915
PotentialHostGenes (PHG)	5.954	5.210

* *adapted from Wicker et al[70] classification.*

Supplementary Table 2. Rose genotypes for resequencing. Sampling site, botanical section, expected ploidy levels and summary of genome wide statistics of the variant calling process

Species	Code	Botanical section	Expected Ploidy	Number of properly paired reads	Number of variants	% of HET variants	% of HOM variants
R. damascena*[1]	DAM	*Gallicanae*	4	52,984,347	10,425,174	53.68	46.32
***R. x hybrida* 'La France'[2]**	FRA	*Modern hybrid*	3	68,630,895	10,757,227	76.53	23.47
R. gallica*[2]	GA	*Gallicanae*	4	52,580,303	10,760,957	51.39	48.61
R. gigantea*[2]	GIG	*Chinenses*	2	53,046,567	7,990,290	57.40	42.60
***R. odorata* 'Hume's Blush'[2]**	HUM	*Chinenses*	2	45,178,271	6,524,466	85.98	14.02
***R. majalis*[1]**	MAJ	*Cinnamomeae*	2	42,780,894	9,274,851	30.29	69.71
***R. moschata*[3]**	MOS	*Synstylae*	2	42,568,752	9,703,825	36.59	63.41
***R. chinensis* 'Mutabilis'[1]**	MUT	*Chinenses*	2	41,820,816	7,971,179	73.19	26.81
***R. pendulina*[2]**	PEN	*Cinnamomeae*	4	62,002,036	10,754,583	42.98	57.02
***R. rugosa*[2]**	RUG	*Cinnamomeae*	2	39,288,390	8,663,148	28.40	71.60
***R. chinensis* 'Old Blush'[1]**	OBHt	*Chinenses*	2	337,061,211	4,731,949	99.87	0.13
***R. chinensis* 'Sanguinea'[1]**	SAN	*Chinenses*	2	43,567,724	6,462,397	69.70	30.30
***R. chinensis* 'Spontanea'[3]**	SPO	*Chinenses*	2	45,991,744	7,378,482	49.54	50.46
R. wichurana*[1]	WIC	*Synstylae*	2	56,855,088	9,897,654	43.34	56.66
***R. arvensis*[4]**	ARV	*Synstylae*	2	41,125,578	9,550,469	34.41	65.59

*[1, 2, 3 or 4] : indicate sampling site, "Ecole Normale Supérieure –Lyon-France", "Lyon Botanical garden" or "Odile Masquelier/La Bonne Maison, Lyon- La Mulatière-France", "jardin expérimental de Colmar, France", respectively. *: sequencing performed at Eurofins Genomics, Ebersberg, Germany. All other lines were sequenced at Genoscope, Evry, France.*

Supplementary Table 3. *Rosa chinensis* type II MADS-box genes

SUBFAMILY	Sublineage	*Arabidopsis* + *Petunia* (B-function only)	*Rosa chinensis***
AGL6		AGL6 (AT2G45650.1); AGL13 (AT3G61120.1)	RcHt_S2597.12 / RcHt_S2662.1
AGL2 (SEP)	SEP3	SEPALLATA3/AGL9) (AT1G24260.1)	RcHt_S139.41 / RcHt_S2321.10*
	SEP1/2	SEPALLATA1 (AGL2) (AT5G15800.1); SEPALLATA2 (AGL4) (AT3G02310.1)	RcHt_S2.15 / RcHt_S1953.16*
	SEP4	SEPALLATA4 (AGL3) (AT2G03710.1)	RcHt_S130.36 / RcHt_S4855.2
AGL12		AGL12 (XAL1) (AT1G71692.1)	RcHt_S1096.17 / RcHt_S3949.2
DEF/AP3	AP3	AP3 (AT3G54340.1); PhDEF (CAA49567.1)	RcHt_S1134.7 / RcHt_S1304.8
	TM6	lost in *Arabidopsis*; PhTM6 (AAS46017.1)	RcHt_S1252.14 / RcHt_S2777.3
PI/GLO		PI (AT5G20240.1); PhGLO1 (AAS46018.1); PhGLO2 (CAA49568.1)	RcHt_S117.34 / RcHt_S950.2 / RcHt_S3308.17; RcHt_S117.36* / RcHt_S3308.15*
Bsister		ABS/TT16/AGL32 (AT5G23260.1)	RcHt_S2091.5* / RcHt_S2425.6*
C-function	AG	AG (AT4G18960.1)	RcHt_S161.30 / RcHt_S1924.4
	PLE	SHP1 (AGL1) (AT3G58780.1); SHP2 (AGL5) (AT2G42830.1)	RcHt_S715.9 / RcHt_S1189.20
D-lineage		STK (AGL11) (AT4G09960.1)	RcHt_S421.24 /RcHt_S2338.3*
AP1/SQUA	euAP1	AP1 (AGL7) (AT1G69120.1); CAL (AGL10) (AT1G26310.1)	RcHt_S6504.1
	euFUL	FUL (AGL8) (AT5G60910.1); AGL79 (AT3G30260.1)	RcHt_S130.33 / RcHt_S2333.1; RcHt_S4605.1 / RcHt_S2.17*
SOC1		AGL20/SOC1 (AT2G45660.1); AGL14 (AT4G11880.1); AGL19 (AT4G22950.1); AGL42 (AT5G62165.1); AGL71 (AT5G51870.1)	RcHt_S165.14 / RcHt_S1161.9; RcHt_S2597.14*/ RcHt_S2662.5*; RcHt_S395.28 / RcHt_S3884.9*
SVP		SVP (AGL22) (AT2G22540.1); AGL24 (AT4G24540.1)	RcHt_S94.8 / RcHt_S1492.6*; RcHt_S77.23 / RcHt_S1869.8; RcHt_S1869.9 / RcHt_S77.24; RcHt_S4043.3 + RcHt_S4043.2*
AGL17		AGL17 (AT2G22630.1); AGL16 (AT3G57230.1); AGL21 (AT4G37940.1); ANR1 (AGL44) (AT2G14210.1)	RcHt_S232.16 / RcHt_S4258.1* + RcHt_S4258.2*; RcHt_S1461.7*+RcHt_S1461.8* ; RcHt_S4683.1* RcHt_S363.14*; RcHt_S279.10*/RcHt_S5487.2*
AGL15		AGL15 (AT5G13790.1); AGL18 (AT3G57390.1)	RcHt_S2472.6 / RcHt_S3453.8*
FLC		FLC (AGL25) (AT5G10140.1); MAF1 FLM (AGL27) (AT1G77080.4); MAF2 (AGL31) (AT5G65050.1); MAF3 FCL3 (AGL70) (AT5G65060.1); MAF4 FCL4 (AGL69) (AT5G65070.1); MAF5 (AGL68) (AT5G65080.1)	-

*: Sequences possibly representing pseudo genes or based on erroneous predictions.

**: Nearly identical sequences, possibly representing different alleles of the same locus, are grouped together and separated by a black slash (/).

Supplementary Table 4. Summary of genomic sequencing data for the homozygous *Rc*HzRDP12.

Library	Insert size	Pair count	Genome overage after read trimming
Overlapping paired end library (2×300bp)	491bp±30	40,827,723	40.5
Nextera mate pair library (2×100bp)	3.3kb±0.6	68,974,420	21.3
Nextera mate pair library (2×100bp)	5.5kb±0.9	61,305,734	19.4
Nextera mate pair library (2×100bp)	8.3kb±1.0	88,477,306	27.8
Nextera mate pair library (2×100bp)	11.6kb±1.1	123,471,460	38.1

Supplementary Table 5. Summary of genomic sequencing data for the heterozygous *R. chinensis* 'Old Blush'

Library	Size	Sequence count	Base count	Coverage
Overlapping paired end library	300 bp	770,553,683	151,319,012,793	135
Sized paired end library	500 bp	167,038,243	33,095,924,262	30
Sized paired end library	600 bp	176,184,201	34,552,203,728	31
Sized paired end library	800 bp	169,370,498	33,077,468,773	30
Mate pair library	3 - 5 Kb	264,810,328	42,665,011,566	38
Mate pair library	5 - 8 Kb	167,453,622	27,427,655,764	24
Mate pair library	8 - 11 Kb	159,141,172	26,049,469,809	23

Supplementary Table 6. Heterozygous genome assembly metrics.

	Contigs	Scaffolds
Number	104,181	15,938
L50	12,925 bp	226,811 bp
L90	2,926 bp	52,670 bp
Total length	746,559,525 bp	882,694,078 bp

Supplementary Table 7. Correspondence between protein-coding genes annotated in *R. chinensis* homozygous and heterozygous assemblies. Each value is the number of allele sets containing a specific number of gene models predicted in the homozygous genome (upper row, in bold), and a specific number of gene models predicted in the heterozygous genome (leftmost column, in bold).

		Number of genes from *Rosa chinensis* homozygous genome in the allele set										
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**≥8**	**≥16**	***Sum***
Number of genes from *Rosa chinensis* heterozygous genome in the allele set	**0**		0	568	100	35	11	7	7	22	2	*752*
	1	1,166	7,813	376	72	18	6	3	1	0	0	*9,455*
	2	333	10,148	355	63	20	5	2	0	1	1	*13,928*
	3	438	547	316	72	15	5	2	0	1	0	*1,396*
	4	141	156	156	66	19	5	3	2	1	0	*549*
	5	52	63	72	47	14	6	3	2	0	0	*259*
	6	30	58	42	45	16	6	3	3	0	0	*203*
	7	7	30	20	20	19	14	6	2	2	0	*120*
	≥8	5	44	44	40	43	41	32	23	53	4	*329*
	≥16	0	10	12	17	16	14	15	18	107	87	*296*
	Sum	*5,172*	*18,869*	*1,961*	*542*	*215*	*113*	*76*	*58*	*187*	*94*	*27,287*

Supplementary Table 8. Primers used for real-time quantitative RT-PCR of miR156 and rRNA 5.8S

Mir156_RT	GTTGGCTCTGGTGCAGGGTCCGAGGTATTCGCACCAGAGCCAACGTGCTC
5.8S_RT	TTGTGACACCCAGGCAGACGTGCCCTCG
Mir156_R	GTGCAGGGTCCGAGGT
mir156_F	GTGTTTTTGGTGACAGAAGAGAGT
5.8S_F	CGGCAACGGATATCTCGG
5.8S_R	TGTGACACCCAGGCAGACG

Supplementary Table 9. Primers used for real-time quantitative RT-PCR

ANS_R	AGCGCGACTTGTCCATTTG	RcHm4g0430121_R	AGGACTGTTCTTGTGGCCTT
ANS_F	GTATCTTGGTTGCTAGCCCC	RcHm4g0430121_F	CGATTAGAGCAAGACGGGGT
CHSa_R	CCGAGTATGGCAACATGTCT	RcHm4g0437871_R	ACAGGAATTATGCAGTGACACT
CHSb_F	CCCAAATAGAACACCCCACTCTAG	RcHm4g0437871_F	CGTTGGGATATTGGGTTTGGT
DFR_R	AAGTGAGTCGCCGCCTTT	GAPDH_R	GGATCGATCACATCGACAGA
DFR_F	TCCTAGACCGCGGCTACA	GAPDH_F	GGTCAAGGTCATTGCTTGGT
F3'H_R	GAAGGAGGAAAGCTCACCGA	TUB_R	AGCATGAAATGGATCCTTGG
F3'H_F	CTATTGCCCATTCCACCGTG	TUB_F	ATTGAGCGTCCCACCTACAC
FLS_R	TGCCCTAGTCATCCACATTG	RhTCTP-R2	CTTGGTTGCTCCCTCAATGT
FLS_F	CGTCTTGTCTTTGCTCACTGT	RhTCTP-F2	GATGCTGATGAGGGTGTTGA
RcHm3g048020_R	GTTTTGGCCGTCTCTCTTCG	RhMYB10-F	CAAATGGCATCGAATTCCTCACTTA
RcHm3g048020_F	TTCATCTCTCCCAGCCCTTG	RhMYB10-R	CTCAACTTCCTCTTGTTCAAAGCTC
RhGDS-F	TGTCCAACAACTGAAAGAAGAAGTG		
RhGDS-R	GTTTTCCCAAACTTGTTTGAACTGG		