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

Olivier Raymond, Jerome Gouzy, Jérémy Just, Hélène Badouin, Marion Verdenaud, Arnaud Lemainque, Philippe Vergne, Sandrine Moja, Nathalie Choisne, Caroline C. Pont, et al.

▶ **To cite this version:**

Olivier Raymond, Jerome Gouzy, Jérémy Just, Hélène Badouin, Marion Verdenaud, et al.. The Rosa genome provides new insights into the domestication of modern roses. Nature Genetics, 2018, 50 (6), pp.772-777. 10.1038/s41588-018-0110-3 . hal-01798003

HAL Id: hal-01798003
https://amu.hal.science/hal-01798003

Submitted on 27 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

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

nature genetics
OPEN

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]*

Roses have high cultural and economic importance as ornamental plants and in the perfume industry. We report the rose whole-genome sequencing and assembly and resequencing of major genotypes that contributed to rose domestication. We generated a homozygous genotype from a heterozygous diploid modern rose progenitor, *Rosa chinensis* 'Old Blush'. Using single-molecule real-time sequencing and a meta-assembly approach, we obtained one of the most comprehensive plant genomes to date. Diversity analyses highlighted the mosaic origin of 'La France', one of the first hybrids combining the growth vigor of European species and the recurrent blooming of Chinese species. Genomic segments of Chinese ancestry identified new candidate genes for recurrent blooming. Reconstructing regulatory and secondary metabolism pathways allowed us to propose a model of interconnected regulation of scent and flower color. This genome provides a foundation for understanding the mechanisms governing rose traits and should accelerate improvement in roses, Rosaceae and ornamentals.

Roses are among the most commonly cultivated ornamental plants worldwide. They have been cultivated by humans since antiquity, for example, in China. Ornamental features as well as therapeutic and cosmetic value have certainly motivated rose domestication. The genus *Rosa* contains approximately 200 species, more than half of which are polyploid[1]. Roses have undergone extensive reticulate evolution with interspecific hybridization, introgression and polyploidization. Only 8 to 20 rose species are thought to have contributed to the present complex hybrid rose cultivars, namely *Rosa*×*hybrida*[2]. The Chinese rose *R. chinensis* (diploid) was introduced to Europe in the eighteenth century. This species is considered one of the main species that participated in the subsequent extensive process of hybridization with roses from the European, Mediterranean and Middle Eastern (mostly tetraploid) sections (Supplementary Note 1). These crosses gave rise to hybrid tea rose cultivars, which are the parents of the modern roses with extraordinarily diverse traits[3]. Among the breeding traits originating from Chinese roses, the capacity of recurrent flowering as well as color and scent signatures are key[4]. Despite recent progress[5], the lack of a rose genome sequence has hampered the discovery of the molecular and genetic determinants of these traits and of their breeding history.

Owing to natural autoincompatibility and recent interspecific hybridization, all roses have highly heterozygous genomes[6] that are challenging to assemble[7] despite their relatively small size (560 Mb)[8]. To date, attempts to assemble rose genomes with short reads have led to highly fragmented assemblies composed of thousands of scaffolds (83,139 (ref. [9]) and 15,938 (this study)).

[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. [17]Present address: Centre for Plant Sciences, Faculty of Biological Sciences, University of Leeds, Leeds, UK. [18]These authors contributed equally: Olivier Raymond, Jérôme Gouzy, Jérémy Just, Hélène Badouin, Marion Verdenaud. [19]These authors jointly supervised this work: Mohammed Bendahmane, Jérôme Gouzy, Moussa Benhamed, Patrick Wincker. *e-mail: mohammed.bendahmane@ens-lyon.fr

© 2018 Nature America Inc., part of Springer Nature. All rights reserved.

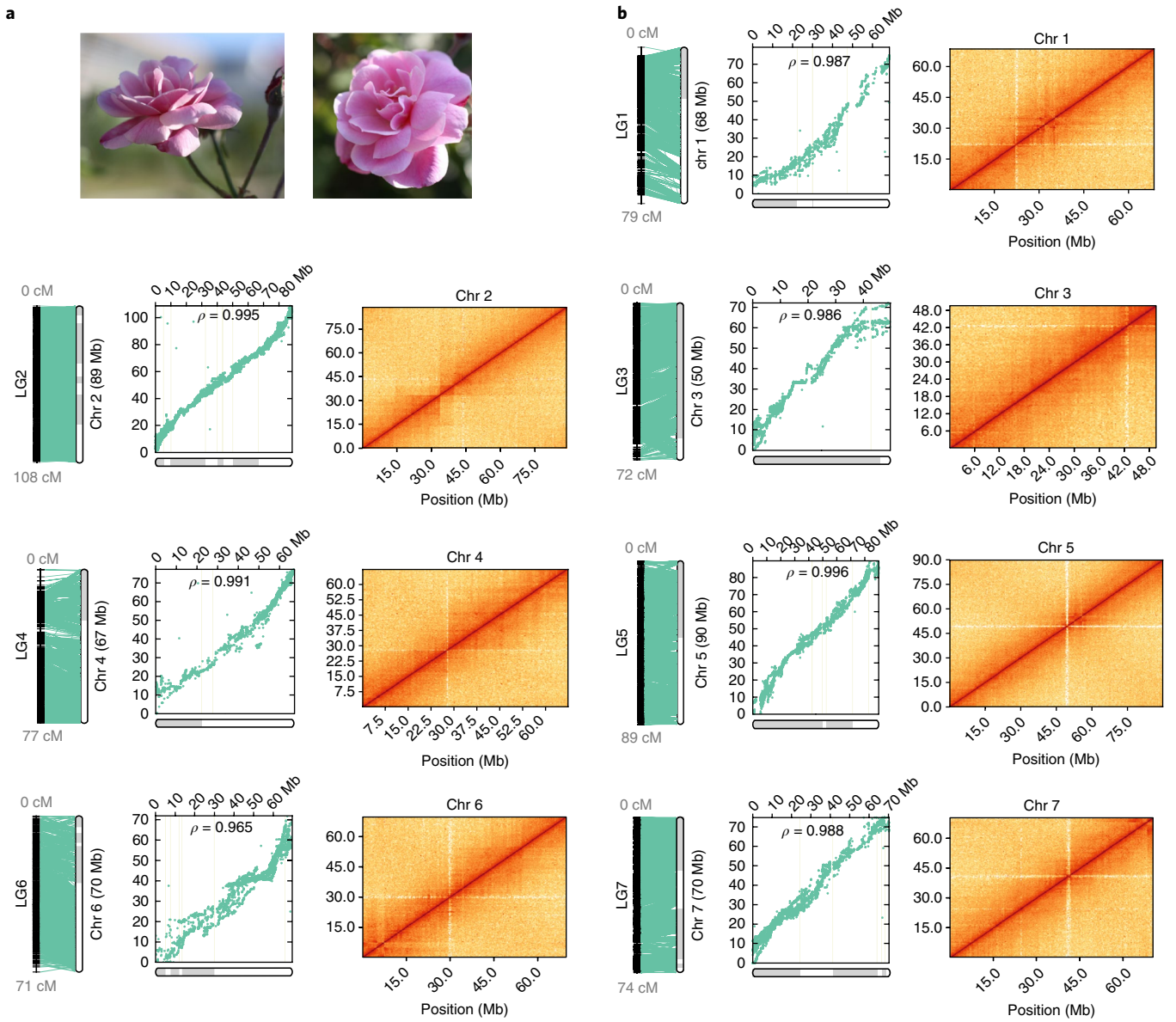

Fig. 1 | Chromosome-level-assembly correlation with genetic map and Hi-C data. a, *R. chinensis* 'Old Blush' mature flowers. **b**, Representations of chromosome connections between the physical positions on the reconstructed chromosome and genetic-map positions (left panels). Scatter plots with dots representing the physical position on the chromosome (Chr) (*x* axis) versus the map position (*y* axis) are shown. Rho (ρ) is the Pearson correlation coefficient (middle panels). A Hi-C intrachromosomal contact map is shown for each chromosome (right panels). The intensity of pixels represents the count of Hi-C links between 400-kb windows on chromosomes on a logarithmic scale. Darker red color indicates higher contact probability. LG, linkage group.

To overcome these bottlenecks to producing a reference genome, we obtained a homozygous genome that we sequenced with long-read sequencing technology. We developed an original in vitro culture protocol combining fine-tuned starvation, cold stress and hormonal treatments to induce *R. chinensis* 'Old Blush' microspores to switch from gametophyte to sporophyte development. This approach allowed microspores to initiate divisions, form homozygous cell clusters and develop embryogenic callus from which homozygous plantlets could be regenerated (Supplementary Note 2 and Supplementary Fig. 1).

The homozygous rose line was sequenced on the PacBio RS II platform. A sequencing coverage of 80× was obtained with 40 single-molecule real-time cells. Preliminary assembly of the rose data with a single assembler generated several hundred contigs, thus illustrating the challenge of assembling plant genomes, even with long-read data[10,11]. A key step in improving the contiguity of the assembly is the detection and the filtering of spurious edges in the graph of overlaps. The assembler CANU implements filter parametrization at the read level, thus leading to more accurate and contiguous assemblies[12]. We developed software called til-r, which implements similar and alternate heuristics to clean the graph of overlaps of the FALCON assembler[13] (Supplementary Fig. 2; URLs). We then used CANU to perform meta-assembly of six complementary raw assemblies generated by CANU and FALCON/TIL-R (Supplementary Note 3; URLs). The final assembly was composed of 82 contigs for an N50 of 24 Mb (where N50 is the contig length, such that 50% of the entire assembly is contained in contigs equal or larger than this value), thus increasing the contiguity metrics of a simple assembly by threefold and demonstrating the power of meta-assembly approaches (Supplementary Fig. 2).

The seven pseudochromosomes were built by integrating 86.4% of the 25,695 markers of the K5 rose high-density genetic map[14].

© 2018 Nature America Inc., part of Springer Nature. All rights reserved.

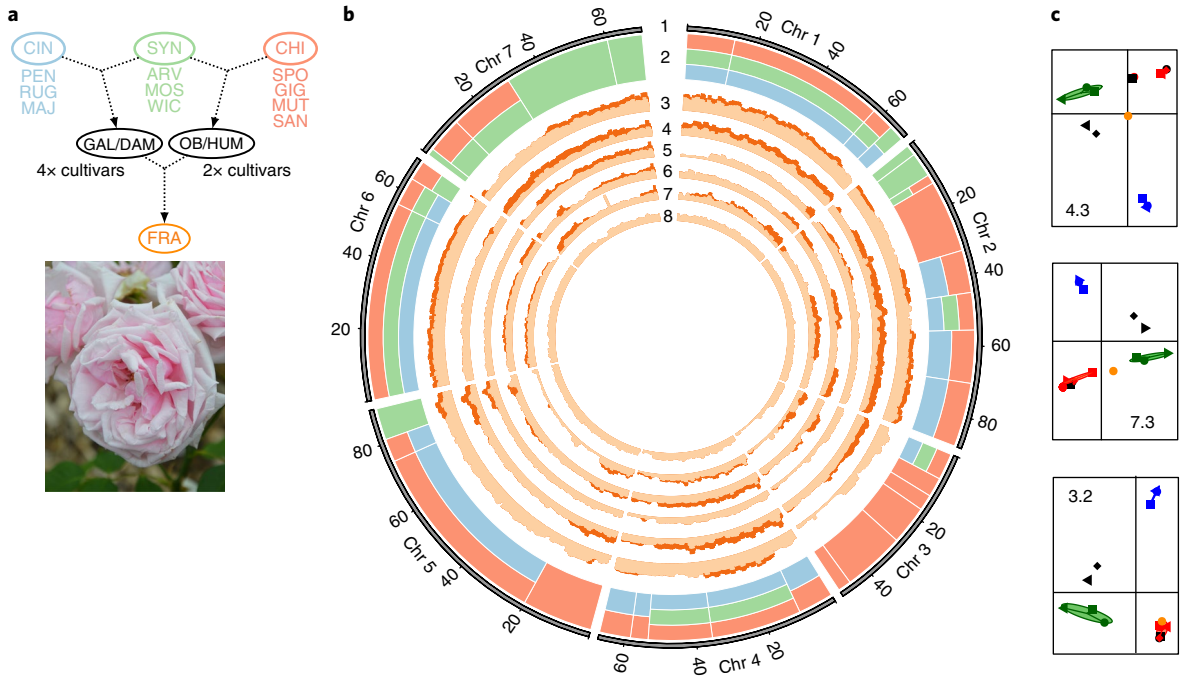

Fig. 2 | Structure of diversity in resequenced genotypes highlights the origin of modern rose cultivars. a, Genealogy of resequenced genotypes. Sections: CIN, Cinnamomeae; SYN, Synstylae; CHI, Chinenses. Genotypes: PEN, *Rosa pendulina*; RUG, *Rosa rugosa*; MAJ, *Rosa majalis*; ARV, *Rosa arvensis*; MOS, *Rosa moschata*; WIC, *R. wichurana*; SPO, *R. chinensis* 'Spontanea'; GIG, *Rosa gigantea*; MUT, *R. chinensis* 'Mutabilis'; SAN, *R. chinensis* 'Sanguinea'; GAL, *Rosa gallica*; DAM, *Rosa damascena*; OB, *R. chinensis* 'Old Blush'; HUM, *R. chinensis* 'Hume's Blush'; FRA, *Rosa* × *hybrida* 'La France' (flower photo). **b**, Genetic structure and variant density. 1, circular representation of pseudomolecules. 2, schematic representation of the contribution of Cinnamomeae, Synstylae and Chinenses sections to 'La France' in 35 chromosomal segments: light red, CHI; light green, SYN; light blue, CIN; multiple bands, mixed origin in the fragment. 3–8, density in heterozygote and homozygote variants (light and dark shades respectively) in 1-Mb sliding windows in 'La France', *R. gigantea*, 'Hume's Blush', 'Mutabilis', 'Sanguinea' and 'Old Blush' heterozygote genotypes, respectively. **c**, Principal component analyses of genetic variation in three illustrative genomic segments. Orange dot, 'La France'; blue, CIN; green, SYN; red, CHI; black, other cultivars. *y* axis, first component; *x* axis, second component. The number indicated in each plot refers to the genomic fragments analyzed (e.g., 4.3 is the third segment of chromosome 4; Supplementary Fig. 6).

A large fraction of the assembly (97.7%, 503 Mb) was oriented with Pearson's correlation coefficients ranging from 0.986 to 0.996, thus illustrating the high congruence between sequence and genetic data. The genome structure and quality were confirmed by mapping of Hi-C chromosomal-contact-map data (Fig. 1 and Supplementary Fig. 3). With its very few remaining gaps and high consistency between genetics and sequence data, the rose genome assembly is one of the most contiguous obtained to date for a plant genome.

The rose genome comprises 36,377 inferred protein-coding genes and 3,971 long noncoding RNAs. Annotation assessment with the Plantae BUSCO v2 dataset[15] identified 96.5% complete gene models. BUSCO analyses using the assembled heterozygous genome of *R. chinensis* 'Old Blush' (Supplementary Note 4) identified 93.5% complete genes (Supplementary Data 1). On the basis of transcriptomic data from pooled tissues, 207 miRNA precursors were predicted. Transposable elements (TEs) spanned 67.9% of the assembly, and 50.6% were long-terminal-repeat retrotransposons (Supplementary Note 5, Supplementary Fig. 4 and Supplementary Table 1). The web portal RchiOBHm-V2 (see URLs) provides access to the reference genome integrating annotations, polymorphisms, transcriptomic data and the first rose epigenome on rose petals (Supplementary Note 6).

Comparative genomic investigation allowed us to assess rose paleohistory within the Rosaceae family (Supplementary Note 7). Conserved gene adjacencies identified an ancestral Rosaceae karyotype consisting of nine protochromosomes with 8,861 protogenes (Supplementary Fig. 5a). Our evolutionary scenario established that the ancestral Rosoideae karyotype of the strawberry and *Rosa* genomes, structured into eight protochromosomes with 13,070 protogenes, was derived from the ancestral Rosaceae karyotype through one ancestral chromosome fission and two fusions. Interestingly, the strawberry genome experienced an extra ancestral chromosome fusion from the ancestral Rosoideae karyotype to reach its modern genome structure, whereas the *Rosa* sp. went through one fission and two fusions, independently of strawberry, to reach its modern genome structure. A phylogeny based on 748 gene sequences showed that *Rosa*, *Fragaria* and *Rubus* diverged within a short timeframe, thus suggesting an evolutionary radiation inside the Rosoideae subfamily (Supplementary Fig. 5b).

To gain insight into the makeup of modern roses, we resequenced representatives of three sections (Synstylae, Chinenses and Cinnamomeae; Supplementary Table 2) that were involved in the domestication and breeding that led to rose hybrid cultivar creation (Supplementary Notes 1 and 8). We observed discrete levels of variant density along the genomes of hybrid cultivars (Fig. 2b) that may reflect different introgression histories. We used the changes in variant density to segment the genome into 35 intervals (2–56 Mb) and studied their genetic structure through principal component analysis (Fig. 2c and Supplementary Fig. 6). We focused on the modern *Rosa* × *hybrida* 'La France', which is considered to be among the first created hybrids combining the growth-vigor traits of European species and the recurrent blooming of Chinese species.

Patterns of diversity along the seven chromosomes showed that the genome of 'La France' is a complex mosaic formed by DNA fragments transmitted by the three ancestral pools of diversity represented in the targeted rose sections (Fig. 2, Supplementary

© 2018 Nature America Inc., part of Springer Nature. All rights reserved.

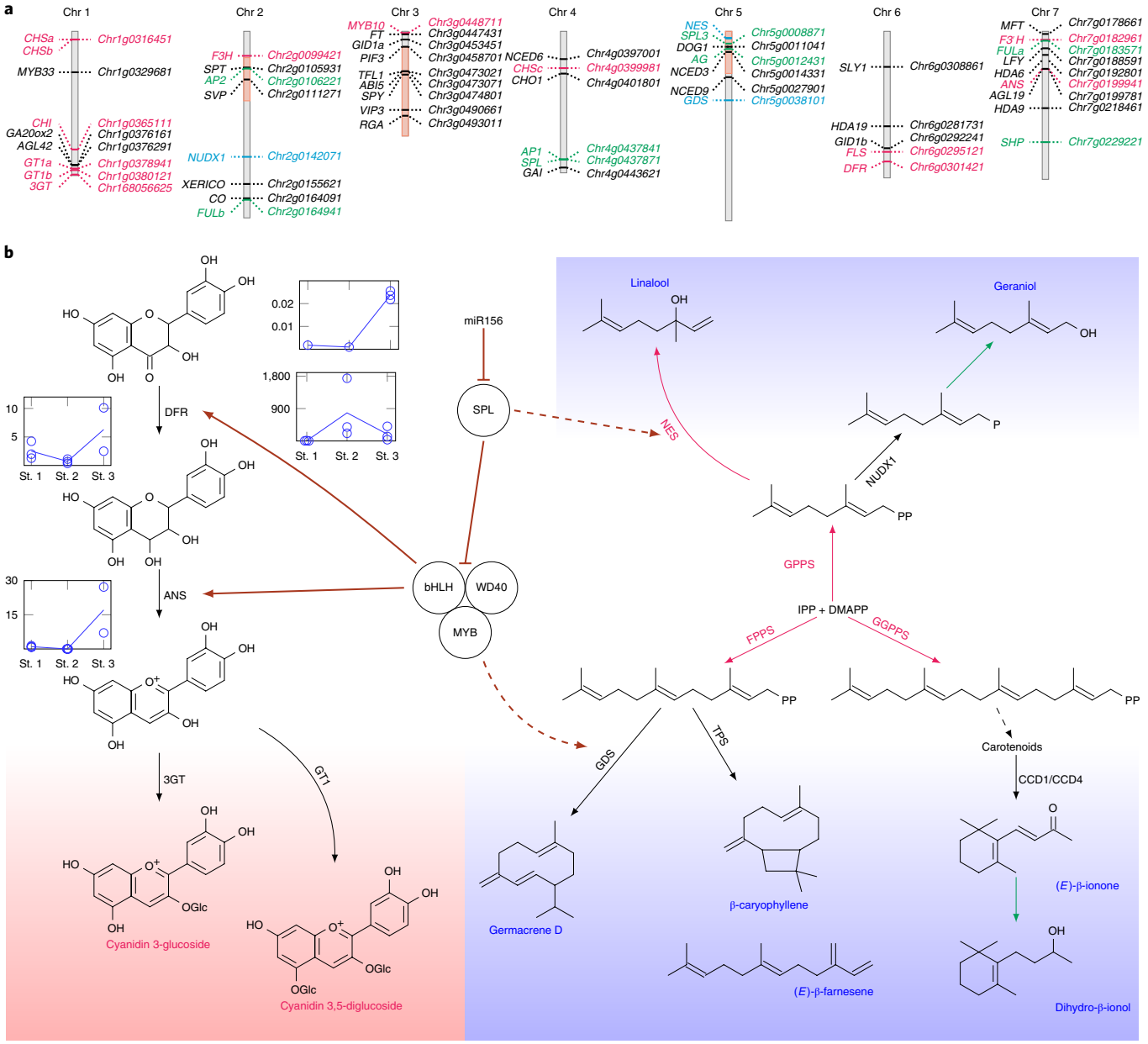

Fig. 3 | Inter-regulatory connections between color biosynthesis and some scent pathways. a, Schematic representation of the rose chromosomes together with the positions of candidate genes potentially affecting anthocyanin pigments, volatile-molecule biosynthesis and flowering. Chromosome segments 2.4, 3.2–3.6 and 5.1 originating only from *R. chinensis* are indicated in light red. Red, anthocyanin-synthesis genes; blue, terpene-biosynthesis genes; black, flowering-time genes; green, development genes. **b**, Schematic representation of interconnections between color (pink background) and scent (blue background) pathways. Gene expression data show the anticorrelation between expression of miR156 and *SPL9* genes during petal development. RT–qPCR was performed on petals harvested at three successive stages: noncolored petals early during development (St. 1); petals at the onset of anthocyanin synthesis (St. 2); and fully colored petals (St. 3). Black arrows, biosynthetic steps reported in rose; red arrows, biosynthetic steps reported in other species but not in rose; green arrows, putative steps with unknown enzymes; dashed black arrow, several enzymatic steps; maroon arrows, gene regulation reported in *Arabidopsis thaliana* but not in rose; dashed maroon arrow, putative gene regulation. IPP, isopentenyl diphosphate; DMAPP, dimethylallyl diphosphate; DFR, dihydroflavonol-4-reductase; ANS, anthocyanidin synthase; 3GT, anthocyanidin 3-*O*-glucosyltransferase; GT1, anthocyanidin 3,5-diglucosyltransferase; GPPS, geranyl diphosphate synthase; FPPS, farnesyl diphosphate synthase; GGPPS, geranylgeranyl diphosphate synthase; GDS, germacrene D synthase; TPS, terpene synthase; NES, linalool/nerolidol synthase; CCD1/CCD4, carotenoid cleavage dioxygenases 1/4; NUDX1, nudix hydrolase 1.

Note 8, Supplementary Fig. 6 and Supplementary Data 2). For example, chromosome 4 haplotypes are structured by a combination of Cinnamomeae, Synstylae and Chinenses genomes, whereas chromosome 7 haplotypes have been transmitted by Synstylae and Chinenses ancestors, without an apparent contribution of Cinnamomeae.

We took advantage of the transmission of genomic bits of Chinenses hybrids to 'La France' to identify new candidate genes potentially involved in recurrent blooming. The insertion of a TE in *TFL1* (*RoKSN*), a repressor of floral transition responsive to activation by gibberellic acid, is considered a major determinant of recurrent blooming[16]. We found that this TE was transmitted

© 2018 Nature America Inc., part of Springer Nature. All rights reserved.

to 'La France' by *R. chinensis* cultivars and thus may participate in its recurrent blooming. A recent segregation analyses of *R. chinensis* 'Old Blush' × *Rosa wichurana* backcross progeny has shown that recurrent blooming probably involves at least a second independent locus[17]. This second locus may have been transmitted to 'La France' by only *R. chinensis* and thus may be located in chromosomal segments such as those originating from the Chinenses section, e.g., segments 2.4 and 5.1 (Fig. 2). On these segments, we identified the putative homologs of the transcription factor *SPT* (segment 2.4, Fig. 3a), which is known to control flowering in *Arabidopsis*[18,19], and of *DOG1* (segment 5.1, Fig. 3a), which is known to modify flowering by acting on miR156 (ref. [20]). These genes are thus additional promising candidates that may determine recurrent blooming in roses.

Roses exhibit a high diversity of flower fragrance and color, of which biochemical and regulatory determinants have been only partially elucidated (Supplementary Note 9 and Supplementary Fig. 7). Data mining of the rose genome combined with in-depth biochemical and molecular analyses of volatile organic compounds permitted identification of at least 22 biosynthetic steps in the terpene pathway that have not been characterized in the rose, two of which have not previously been characterized in other species (Supplementary Note 9 and Supplementary Fig. 7).

To study the relationships between color and scent pathways, we performed biochemical and molecular analyses on cyanidin, whose glucosylated derivatives represent more than 99% of the total anthocyanin pigments[21], and on germacrene D, a volatile organic compound produced in petal cells of *R. chinensis* 'Old Blush' (Supplementary Data 3). Our analyses suggest that coordinated biosynthesis of these two compounds is achieved through the miR156–*SPL9* regulatory module. In *Arabidopsis*, *SPL9* is a repressor of anthocyanin synthesis in the cells of aging plants[22]. miR156 negatively regulates *SPL9* in the cells of young plants, thereby enabling the formation of a MYB–bHLH–WD40 protein complex that activates anthocyanin production[22]. Analysis of this module in the petals of 'Old Blush' indicated that the expression of *SPL9* peaked before maximum *ANTHOCYANIDIN SYNTHASE (ANS)* expression (Supplementary Fig. 8). In fully colored petals, we observed induced expression of miR156, which correlated with downregulation of *SPL9* expression and upregulation of *ANS* expression (Fig. 3b, Supplementary Fig. 8 and Supplementary Fig. 9). The maximum expression of *GDS*, which encodes the enzyme catalyzing germacrene D synthesis, also correlated with miR156 and *ANS* activation and with *SPL9* downregulation (Fig. 3 and Supplementary Fig. 8). This observation, together with a previous demonstration that *ANS* and *GDS* can be activated in rose petals by expression of the *Arabidopsis AtPAP1* MYB transcription factor[23], suggests that the biosynthesis of anthocyanin and germacrene D may be modulated by the miR156–*SPL9* regulatory module, possibly through action on a MYB–bHLH–WD40 complex. Although *PAP1* is not expressed in 'Old Blush' petals, we found that the expression pattern of *RhMYB10*, which has been described as a regulator of the anthocyanin-biosynthetic pathway in *Rosaceae*[24], is compatible with a role in coactivation of the synthesis of cyanidin and germacrene D in petal epidermal cells (Supplementary Fig. 8).

The biosynthesis of terpenes, major scent compounds in roses, has been shown to involve TERPENE SYNTHASE (TPS) proteins, such as NEROLIDOL SYNTHASE (NES)[25]. A search for *TPS* in the rose genome revealed a cluster of *NES* genes on chromosome 5 that has a counterpart in *Fragaria*[26]. These genes were not substantially expressed in rose petals (Supplementary Data 4). In *Arabidopsis*, the expression of some *TPS* is activated by *SPL9* (ref. [27]). In rose petals, the downregulation of *SPL9* through activation of miR156 (Fig. 3b and Supplementary Fig. 8) might explain the absence of expression of *NES* genes and probably explains why they do not participate in the production of some terpenes in rose flowers. Our data provide hints as to why alternative routes to produce terpenes, such as the one involving *NUDX1* (ref. [28]), have been used in rose flowers.

Here, we propose that the miR156–*SPL9* regulatory hub orchestrates the coordination of production of both colored anthocyanins and certain terpenes, by permitting the complexation of preexisting MYB–bHLH–WD40 proteins, which in turn modulate different components of both pathways (Fig. 3). Therefore, anthocyanin synthesis in rose flowers may be linked to the production of some volatile compounds, thus providing a regulatory explanation for the evolution of nonstandard terpene-biosynthesis pathways. Moreover, this co-regulation may hinder the combination of pigmentation and specific scents in rose hybrids.

The very high-quality rose genome sequence reported in this study, combined with an expert annotation of the main pathways of interest for the rose (Supplementary Notes 9–13, Supplementary Figs. 7–23, Supplementary Table 3 and Supplementary Data 5–10), provides new insights into the genome dynamics of this woody ornamental and offer a basis to disentangle the seemingly mandatory trait associations or exclusions. Furthermore, access to candidate genes, such as those involved in abscisic acid synthesis and signaling, paves the way for improving rose quality with better water-use efficiency and increased vase life. Breeding for other characteristics such as increased resistance to pathogens should also benefit from these data and may lead to decreased use of pesticides.

URLs. Genome browser and genomic resources, https://lipm-browsers.toulouse.inra.fr/pub/RchiOBHm-V2/; MetExplore, https://metexplore.toulouse.inra.fr/metexplore2/?idBioSource=5104/; EuGene plant pipeline, http://eugene.toulouse.inra.fr/Downloads/egnep-Linux-x86_64.1.4.tar.gz; tbl2asn2, https://www.ncbi.nlm.nih.gov/genbank/tbl2asn2/; REPET, https://urgi.versailles.inra.fr/Tools/REPET/; miRanda, http://www.microrna.org/; til-r, http://lipm-bioinfo.toulouse.inra.fr/download/til-r/.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0110-3.

Received: 1 December 2017; Accepted: 14 March 2018;
Published online: 30 April 2018

References

1. Fougère-Danezan, M., Joly, S., Bruneau, A., Gao, X. F. & Zhang, L. B. Phylogeny and biogeography of wild roses with specific attention to polyploids. *Ann. Bot.* **115**, 275–291 (2015).
2. De Vries, D. P. & Dubois, L. Rose breeding: past, present, prospects. *Acta Hortic.* **424**, 241–248 (1996).
3. 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. *Theor. Appl. Genet.* **102**, 398–404 (2001).
4. Hurst, C. C. Notes on the origin and evolution of our garden roses. *J. R. Hort. Soc.* **66**, 73–82 (1941).
5. Bendahmane, M., Dubois, A., Raymond, O. & Bris, M. L. Genetics and genomics of flower initiation and development in roses. *J. Exp. Bot.* **64**, 847–857 (2013).
6. Esselink, G. D., Smulders, M. J. & Vosman, B. Identification of cut rose (*Rosa hybrida*) and rootstock varieties using robust sequence tagged microsatellite site markers. *Theor. Appl. Genet.* **106**, 277–286 (2003).
7. Zharkikh, A. et al. Sequencing and assembly of highly heterozygous genome of *Vitis vinifera* L. cv Pinot Noir: problems and solutions. *J. Biotechnol.* **136**, 38–43 (2008).
8. Yokoya, K., Roberts, A. V., Mottley, J., Lewis, R. & Brandham, P. E. Nuclear DNA amounts in roses. *Ann. Bot.* **85**, 557–561 (2000).
9. Nakamura, N. et al. Genome structure of *Rosa multiflora*, a wild ancestor of cultivated roses. *DNA Res.* https://doi.org/10.1093/dnares/dsx042 (2017).
10. Badouin, H. et al. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* **546**, 148–152 (2017).
11. VanBuren, R. et al. Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature* **527**, 508–511 (2015).

© 2018 Nature America Inc., part of Springer Nature. All rights reserved.

12. 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).
13. Chin, C. S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
14. Bourke, P. M. et al. Partial preferential chromosome pairing is genotype dependent in tetraploid rose. *Plant J.* **90**, 330–343 (2017).
15. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
16. Iwata, H. et al. The TFL1 homologue KSN is a regulator of continuous flowering in rose and strawberry. *Plant J.* **69**, 116–125 (2012).
17. Li, S. et al. Inheritance of perpetual blooming in *Rosa chinensis* 'Old Blush'. *Hortic. Plant J.* **1**, 108–112 (2015).
18. Mouradov, A., Cremer, F. & Coupland, G. Control of flowering time: interacting pathways as a basis for diversity. *Plant Cell* **14** Suppl, S111–S130 (2002).
19. Vaistij, F. E. et al. Differential control of seed primary dormancy in *Arabidopsis* ecotypes by the transcription factor SPATULA. *Proc. Natl Acad. Sci. USA* **110**, 10866–10871 (2013).
20. Huo, H., Wei, S. & Bradford, K. J. DELAY OF GERMINATION1 (DOG1) regulates both seed dormancy and flowering time through microRNA pathways. *Proc. Natl. Acad. Sci. USA* **113**, E2199–E2206 (2016).
21. Han, Y. et al. Comparative RNA-seq analysis of transcriptome dynamics during petal development in *Rosa chinensis*. *Sci. Rep.* **7**, 43382 (2017).
22. 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**, 1512–1522 (2011).
23. Zvi, M. M. et al. PAP1 transcription factor enhances production of phenylpropanoid and terpenoid scent compounds in rose flowers. *New Phytol.* **195**, 335–345 (2012).
24. 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).
25. Aharoni, A. et al. Gain and loss of fruit flavor compounds produced by wild and cultivated strawberry species. *Plant Cell* **16**, 3110–3131 (2004).
26. Shulaev, V. et al. The genome of woodland strawberry (*Fragaria vesca*). *Nat. Genet.* **43**, 109–116 (2011).
27. Yu, Z. X. et al. Progressive regulation of sesquiterpene biosynthesis in *Arabidopsis* and Patchouli (*Pogostemon cablin*) by the miR156-targeted SPL transcription factors. *Mol. Plant* **8**, 98–110 (2015).
28. Magnard, J. L. et al. Biosynthesis of monoterpene scent compounds in roses. *Science* **349**, 81–83 (2015).

Acknowledgements

We thank J. Thomas, T. Goujon and C. Bendahmane for critical reading of the manuscript. We thank A. Meilland for helpful discussions. We thank A. Lacroix, P. Bolland and J. Berger (ENS de Lyon, France) for plant handling. We thank the Lyon Botanical Garden–France and the rose garden La Bonne Maison O. Masquelier, Lyon, France, for providing plant material. We thank the Genotoul bioinformatics platform Toulouse Midi-Pyrenées for providing help and computing resources and L. Taulelle and E. Quemener (ENS de Lyon) for assistance with computing. We gratefully acknowledge support from the Pôle Scientifique de Modélisation Numérique of the ENS de Lyon for computing resources. We thank the epigenomic platform of the IPS2-University Paris-Sud-Orsay France. We thank the platforms 'AniRA-Cytometry' and 'Analyse Génétiques et Cellulaire' of the IFR BioScience Lyon (UMS3444/US8) for HRM and flow cytometry experiments. The Get-Plage platform was supported by the GET-PACBIO program (programme operationnel FEDER-FSE MIDI-PYRENEES ET GARONNE 2014-2020). This work was supported by funds from the French National Institute of Agronomic Research (INRA); the program Fonds Recherche of Ecole Normale Supérieure-Lyon-France to M. Bendahmane and O.R.; the Genoscope to P.W.; the French National Research Agency programs DODO (ANR-16CE20-0024-03 to M. Bendahmane and M. Vandenbussche) and AUXIFLO (ANR-12-BSV6-0005 to T.V); and the European Research Council (ERC-SEXYPARTH) and the Labex Saclay Plant Sciences–SPS (ANR-10-LABX-0040-SPS) to A. Bendahmane.

Author contributions

O.R. and C.B. performed DNA extraction. P. Vergne and P. Villand produced the rose homozygous line. C.L.-R. and M. Bendahmane performed PacBio sequencing data production. J. Szécsi performed flow cytometry experiments. P.H. provided *Rosa* material. M. Verdenaud, D.L., M. Benhamed and M.P. performed epigenome analysis. O.R., X.F., S.-H.Y., A.D., M.L.B. and M. Bendahmane performed DNA/RNA sample collection and data production. J.J., M. Verdenaud, M.L., L.F. and O.R. performed RNA-seq and analyses of gene expression. M. Benhamed, M. Verdenaud, C.L., A. Boualem and A. Bendahmane performed chromosome conformation capture Hi-C. M. Benhamed and M. Verdenaud integrated the assembly and genetic map to build pseudochromosomes. J.G. developed bioinformatics tools and assembled the PacBio homozygous genome. M. Benhamed and M. Verdenaud validated the assembly with Hi-C and genetic data. P.W. and A. Lemainque performed Illumina sequencing. A. Boualem and A. Bendahmane provided resequencing sequencing data. A.C., J.-M.A., A.M., K.L. and P.W. assembled the heterozygous rose Illumina sequencing data. N.C., H.Q. and J.J. conducted repetitive DNA analysis. J.G., N.C. and J.J. annotated protein-coding genes, TEs and miRNAs. J.G., H.B., J.J. and L.C. performed bioinformatics analyses. S.C. built the *Rosa* web portal. J. Salse and C.P. conducted paleoevolution analyses. J.J. and O.R. conducted miRNA analyses. A. Larrieu, T.V. and J.J. performed integrated analyses on auxin genes. S.M., J.-C.C. and S.B. performed gas chromatography–mass spectrometry analyses of scent compounds. S.M., J.-C.C., S.B., O.R. and J.J. performed integrated analyses on scent genes. O.R., L.P., F.P., L.F. and M. Verdenaud performed integrated analyses on flowering genes. M. Vandenbussche performed integrated analyses on MADS transcription-factor genes. O.R., L.F., J.J., L.P., J. Szécsi. and V.B. performed integrated analyses on color genes. M.L.-B., B.G. and Y.L. performed integrated analyses on meiosis genes. H.B., O.R. J.G., J.J. and F.P. performed diversity analysis. M. Bendahmane and J.G. coordinated the rose genome consortium. M. Bendahmane, O.R., H.B. and J.G. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0110-3.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© 2018 Nature America Inc., part of Springer Nature. All rights reserved.

Methods

Production of homozygous rose line derived from heterozygous *R. chinensis* 'Old Blush'. Flower buds were harvested from *R. chinensis* 'Old Blush' plants when most microspores were at the mid–late uninucleate/early bicellular development stages (Supplementary Fig. 1). Microspores were aseptically isolated from anthers, suspended in starvation medium and pretreated at 4 °C in darkness for 21 d. Approximately 160,000 microspores were suspended in AT12 medium corresponding to AT3 medium[29] supplemented with 4.5 µM 2,4-D and 0.44 µM BAP, pH 5.8, and then incubated at 25 °C in the dark. Developing microcalli (~0.5 mm diameter) were observed after approximately 11 weeks and were then subcultured individually under the same conditions (Supplementary Note 2). Developed calli were then plated onto solid MS salt medium complemented with B5 vitamins, 30 g/L sucrose, 2.5 mM MES, 4.5 µM 2,4-D, 0.44 µM BAP and 6.5 g/L VitroAgar (Kalys Biotechnologie), pH 5.8. A callus that displayed somatic embryos (designated *Rc*HzRDP12; Supplementary Fig. 1g) was selected. The homozygosity status and ploidy level of this callus were confirmed by DNA genotyping and fluorescence-activated cell-sorting analysis, respectively, as previously described[30].

Sample preparation and sequencing. High-quality nuclear DNA was prepared from *Rc*HzRDP12 homozygous callus propagated on callus-maintenance medium (Supplementary Note 2) as previously described[31] with the following modifications. Ten percent fresh weight of PVP40 was added to callus cells that had been ground in liquid nitrogen. Purified nuclei pellets were processed with a Qiagen DNeasy Plant kit (Qiagen). DNA integrity was verified via gel electrophoresis (0.7% agarose), and total DNA was quantified through fluorometry with Picogreen (Applied Biosystems/Life Technologies).

To sequence the *R. chinensis* 'Old Blush' genome, we used in vitro–cultured plants obtained through adventitious shoot organogenesis from type 1 somatic embryo (*Rc*OBType1), as previously described[32]. Axenic in vitro *R. chinensis* 'Old Blush' plantlets were ground in liquid nitrogen, and nuclei were purified as previously described[31]. Nuclei pellets were then processed with a Qiagen DNeasy Plant kit (Qiagen), according to the protocol provided by the supplier.

High-quality DNA was extracted from leaf samples of *Rosa* species and cultivars grown at ENS-Lyon, at the Lyon botanical garden, in the rose garden 'La Bonne Maison, O. Masquier, Lyon, France' or in the rose garden 'Jardin Expérimental de Colmar, France' (Supplementary Note 8).

DNA integrity was verified by gel electrophoresis (0.7% agarose), and DNA was then quantified by fluorometry with Picogreen (Applied Biosystems/Life Technologies).

Paired-end-sequencing DNA libraries were constructed with Illumina's TruSeq DNA LT kit according to the manufacturer's recommendations (Supplementary Tables 4 and 5). The distributions of DNA-fragment lengths in the libraries were verified with Agilent BioAnalyzer High Sensitivity DNA chip assays. Whole-genome sequencing of *R. chinensis* 'Old Blush' was performed on an Illumina HiSeq 2000 instrument. Sequences from paired-end and mate-pair reads of the multiple libraries were assembled in ALLPathsLG software[33] (Supplementary Table 6).

Three-dimensional proximity information obtained by chromosome conformation capture sequencing (Hi-C). Leaf tissues were fixed in 1% (vol/vol) formaldehyde and were then used for preparation of two independent in situ Hi-C libraries. Nuclei extraction, nuclei permeabilization, chromatin digestion and proximity-ligation treatments were performed essentially as previously described[34]. DpnII was used as a restriction enzyme. The recovery of Hi-C DNA and subsequent DNA manipulations were performed as previously described[35]. Libraries were sequenced on an Illumina NextSeq instrument with 2 × 75-bp reads. Hi-C libraries were independently analyzed in HiC-Pro pipeline (default parameters and LIGATION_SITE = GATCGATC[36]). Valid ligation products from each library were merged for interaction-matrix construction. The genome was divided into bins of equal size, and the number of contacts was determined between each pair of reported bins. Finally, contact maps were plotted in HiCPlotter software[37].

Genome assembly. The program til-r was developed to implement heuristics aiming at filtering the graph of overlap generated by FALCON (Supplementary Note 3). A meta-assembly combining two CANU and four FALCON assemblies was generated in CANU 1.4 (Supplementary Fig. 2 and Supplementary Note 3).

Pseudomolecule building. Pseudomolecules were built by anchoring the 82 contigs to the K5 SNP genetic linkage map[14] in ALLMAPS software[38]. Four chimeric breakpoints were identified and corrected by identifying the primary contigs in which the problematic regions were not merged. Three chimeric breakpoints were absent in CANU assemblies, and the fourth was absent in all primary assemblies. Finally, ALLMAPS was applied on the corrected meta-assembly, thus enabling building of seven pseudomolecules corresponding to the rose haploid chromosome number by anchoring and orienting 97.7% of the contigs (503 Mb) based on 86.4% of the genetic markers. The final assembly consists of seven pseudochromosomes and the mitochondrial and chloroplast genomes plus 46 unanchored contigs spanning 11.2 Mb (Supplementary Fig. 2a).

The genome was first polished in quiver[39] with stringent alignment cutoffs (--minLength 3000 --maxHits 1). Then, a run of pilon[40] (version 1.21, --mindepth 30 --fix bases) with homozygous 'Old Blush' Illumina paired-end reads edited 7,444 SNPs, 107,249 small insertions and 33 small deletions. The final genome assembly is composed of 515,588,973 nt including the 3,300 'N' for the 33 gaps, seven of which represent centromeres. Biological centromeres were located by identifying tandem repeats in TRF software[41], selecting patterns of an over-represented length in the genome, assembling them in contigs and visually inspecting their distribution along the pseudomolecules (Supplementary Note 3).

Localization of putative crossovers and segmental conservation between genotypes. Identification of putative loci of crossovers was performed by mapping Illumina reads from the heterozygous genome (five distinct libraries) on the constructed pseudochromosomes in BWA software[42] and by counting pairs in which only one read had a match, in 10-kb-long windows. We observed 50 windows with over-represented one-end-mapped pairs in at least two libraries and kept them as candidate crossover loci (Supplementary Fig. 12, yellow frame). To confirm them, when possible, we used the sequence conservation with genotypes related to the inferred parents of 'Old Blush' (Supplementary Fig. 12, red plots; Supplementary Note 4.2).

Annotation of protein-coding genes and lncRNAs. Gene models were predicted with a fully automated and parallelized pipeline, egn-ep (see URLs), that carries out probabilistic sequence model training, genome masking, transcript and protein alignment computation and integrative gene modeling in EuGene software[43] (release 4.2a). The configuration of the egn-ep pipeline is detailed in Supplementary Note 5. The inferred mRNAs were assessed in BUSCO v2 (ref. [15]), which found 1,389 complete, 23 fragmented and 28 missing gene models (96.5%, 1.6% and 1.9% respectively). 36,377 genes were retained after the removal of annotated repeated elements (described below). The correspondence between gene models in homozygous and heterozygous annotations was established on the basis of best reciprocal hits (Supplementary Table 7 and Supplementary Data 1).

Functional annotation of protein-coding genes. The protocol described by Schläpfer et al.[44] was used to annotate enzymes and build the metabolic network. Two cutoffs were modified to increase stringency: the BLAST e-value cutoff was lowered to 10^{-5}, and the pathway-prediction score was set to 0.3 in pathway-tools. Nineteen pathways considered to be false positives were removed. A MetExplore instance[45] is available to visualize the network (see URLs).

Protein-coding genes were annotated through integration of five sources, depending on their expected accuracy. Priorities were successively given to (i) a search of reciprocal best hits with the 218 Rosaceae proteins tagged as 'reviewed' in the UniProt database (90% span, 80% identity)[46], (ii) the description of the 8,512 previously annotated enzymes, (iii) transcription factors and kinases identified (2,414 and 1,885 respectively) by ITAK[47], (iv) the 3,954 transcription factors identified by PlantTFCat[48] and (v) the InterPro analysis matching 31,853 proteins[49]. Finally, the annotations were tested and edited when needed to follow consistency rules defined by GenBank (see URLs).

De novo transposable-element and repeat annotation. The pseudochromosomes were deconstructed into 'virtual' contigs by removal of stretches of >11 undefined bases (Ns) to exclude gaps. We generated 2,742 virtual contigs with an N50 of 22 Mb for a total length of 515 Mb. The TE*denovo* pipeline[50,51] from the REPET package v2.5 (see URLs) was used to detect TEs in these contigs and to build a consensus sequence for each TE family with a minimum of five sequences per group. A library was generated containing 28,545 consensus sequences, classified according to structural and functional features (similarities with characterized TEs from the RepBase database v21.01 (ref. [52]) and domains from Pfam27.0). After removal of redundancy and filtering consensus sequences classified as satellites (labeled SSR) and unclassified consensus sequences constructed with fewer than ten copies in the genome, a library of 8,226 consensus sequences was used to annotate TE copies in the whole homozygote genome with the TEannot pipeline with default parameters[53]. To refine TE annotation, consensus sequences showing no full-length fragments (i.e., fragments covering more than 95% of the consensus sequence) in the genome were filtered out, and a subset of 3,933 consensus sequences was used to run a second TEannot iteration. After a manual curation step to reclassify some consensus sequences, the final annotation files were renamed with this new classification, and this library was used to annotate the heterozygote genome (15,938 scaffolds for a total length without Ns of 746 Mb) with the TEannot pipeline. Consensus sequences classified as potential host genes bearing Pfam domains were manually curated and removed from the TE set (453 consensus sequences).

Annotation of miRNA precursors and mature miRNAs. To identify *R. chinensis* miRNA genes, an RNA library was constructed with mixed RNAs from pooled organs. After adaptor cleaning and removal of rRNA/tRNA-related sequences, we identified 38 million putative small RNAs displaying a size distribution ranging between 20 and 25 nt, with two peaks at 21 nt (17 million) and 24 nt (11.8 million). Genome-wide annotation of miRNA precursors was performed with an updated

© 2018 Nature America Inc., part of Springer Nature. All rights reserved.

version of the pipeline described by Formey et al.[54], which was modified to integrate stringent criteria proposed by miRBase (for example, expression of both mature 5p and 3p miRNAs)[55]. A total of 207 miRNA precursor loci were predicted to correspond to 636 expressed mature precursors (328 5p and 308 3p). miRNA targets were predicted with miRanda v3.0 (see URLs). Known mature miRNAs not found by the automatic and stringent process were annotated with blastn.

Genetic structure and genome segmentation. Illumina data mapping and SNP calling were performed as described in Supplementary Note 8. The number of homozygote and heterozygote variants in sliding windows of 1 Mb was computed on genic SNPs for each genotype, with functions of the bedtools suite (bedtools makewindows, bedtools intersect and bedtools groupby)[56]. To compute the density of variants per window, the number of variants was divided by the number of informative sites (mapping coverage between 5 and 60 for the 14 resequenced species and between 50 and 300 for the heterozygote Old Blush genotype). We use the term variants in tetraploid species to refer both to allelic differences and to differences between homeologs (i.e., between genes of different subgenomes). Owing to vegetative multiplication of rose cultivars, limited recombination has occurred after hybridization, and the size of introgressed fragments should be large. If the genomes or subgenomes involved in hybridization events have different distances with respect to the reference genome, genomic regions with different introgression histories should display different levels of variant density in resequenced hybrid cultivars. We used the changes in variant density in the genotypes FRA, GIG, HUM, MUT and SAN to segment the genome into 35 intervals (ranging from 2 to 56 Mb). The genomic boundaries were defined as the start of the windows corresponding to the inflexion points in density files. For each of the 35 genome segments, the genetic structure was inferred on biallelic SNPs with no missing data and not overlapping with repeat elements. Principal component analyses[57] were performed with the glPCA function of the adegenet package (version 2.0.1)[58]. Axes 1 and 2 of the PCA explained a significant proportion of the variance (29.29% to 40.53% and 12.07% to 19.89%, respectively). Therefore, we present only the analyses of these two axes.

Rose and Rosaceae paleogenomics. Two parameters were defined as previously described[59] to increase the stringency and significance of BLAST sequence alignment by parsing BLAST results and rebuilding high-scoring pairs or pairwise sequence alignments to identify accurate paralogous and orthologous relationships between *Rosa* (7 chromosomes, 49,767 genes), apricot (8 chromosomes, 31,390 genes), peach (8 chromosomes, 27,864 genes), apple (17 chromosomes, 63,514 genes), pear (17 chromosomes, 42,812 genes) and strawberry (7 chromosomes, 32,831 genes). From the previous orthologous and paralogous relationships, ancestral karyotypes were reconstructed as defined by Salse[59], such that the ancestral genome is a 'median' or 'intermediate' genome consisting of a clean reference-gene order common to the extant species investigated.

Biochemical analyses of scent composition in roses. Volatile compounds were extracted with hexane from petals and stamens of roses of the different genotypes, mainly as previously described[28] (Supplementary Note 9). Camphor was used as an internal standard to estimate compound quantities. Hexane sample fractions were analyzed with a gas chromatograph coupled to an electron ionization mass spectrometer detector (Agilent 6850) operated under an ion-source temperature of 230 °C, a trap emission current of 35 µA and a 70-eV ionization energy. All experiments were performed at least twice. Chromatographs were analyzed in Agilent Data Analysis software, and the volatile substances were identified by screening the WILEY 275, NIST 08 and CNRS libraries to compare MS spectra. The Kovats retention index of each substance was calculated with data of the injection of a homologous set of *n*-alkane (C_8–C_{20}) according to the Kovats formula[60]. Mass-spectra similarities together with Kovats-retention-index values were then used for compound identification. Concentrations were calculated through comparison of the camphor area as the internal standard.

ChIP–seq assays. Petals were collected from *R. chinensis* 'Old Blush' and fixed in 1% (vol/vol) formaldehyde. ChIP assays were performed with anti-H3K9ac (Millipore, 07-352) or anti-H3K27me3 (Millipore, 07-449) according to a procedure adapted from Veluchamy et al.[61]. Library quality was assessed with an Agilent 2100 Bioanalyzer (Agilent), and the libraries were subjected to high-throughput sequencing on an Illumina NextSeq 500 instrument. After trimming, reads were aligned to the *R. chinensis* genome in bowtie2 (ref. [62]) with a maximum mismatch of 1 bp and unique mapping reported. To determine the target regions of H3K9ac ChIP–seq, model-based analysis of ChIP–seq (MACS2)[63] was used. Detection of H3K27me3-modified regions was performed with SICER[64]. HOMER[65] was used to annotate H3K9ac peaks with nearby genes if peaks were located in windows −2 kb to +1 kb around the gene TSS. For H3K27me3 peaks, bedtools intersect[56] was used, and only genes that overlapped with this specific modification were kept. Clustering of H3K9ac and H3K27me3 peaks was performed with SeqMINER[66]. Rstudio, Circos[67] and NGSplot[68] were used for graphic representation of histone modifications.

RNA preparation and qPCR analyses. Total RNA and small RNAs were prepared from petals at three developmental stages: noncolored petals early during development (closed bud; stage 1); petals at the onset of anthocyanin synthesis (closed bud; stage 2); and fully colored petals with maximum anthocyanin content (bud opening; stage 3). Total RNA was prepared as previously described[69]. One microgram of RNA was used in reverse-transcription assays, and qPCR was performed as previously described[70] with gene-specific primers (Supplementary Note 10 and Supplementary Tables 8 and 9). Small RNAs were extracted with a Macherey-Nagel NucleoSpin miRNA kit. Contaminating DNA was removed with an Ambion DNA-free kit. RNA concentrations were measured with a NanoDrop ND-1000 Micro-Volume spectrophotometer (NanoDrop Technologies) before and after DNase treatment. Small-RNA quantification was performed with stem-loop RT–PCR as previously described[71]. Reverse transcription was performed with a RevertAid kit (Thermo Fisher Scientific). Primers specific to 5.8S rRNA or stem-loop RT-primer for miR156 (Supplementary Note 10 and Supplementary Table 8) were used. 5.8S rRNA and miR156 expression were quantified with a QuantStudio 6 Flex Real-Time PCR 384 instrument (Applied Biosystems) with a Fast SYBR Green Master Mix kit (Roche Diagnostic) and specific primers (Supplementary Note 10). Data were collected for three independent biological replicates.

Code availability. Source code (in C) and linux binaries of the til-r software are available at http://lipm-bioinfo.toulouse.inra.fr/download/til-r/ under the GPL license.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary.

Data availability. The *R. chinensis* 'Old Blush' homozygous genome has been deposited in DDBJ/ENA/GenBank under accession number PDCK00000000. PacBio raw data have been deposited in the Sequence Read Archive (SRA) under study accession number SRP119907. The *R. chinensis* 'Old Blush' heterozygous genome has been deposited under BioProject accession number PRJEB24406.

Resequencing sequence reads have been deposited in the SRA under study accession number SRP119986.

Hi-C data have been deposited under SRA accession numbers SRR6189546 and SRR6189547, and ChIP–seq data have been deposited under SRA accession numbers SRR6167310, SRR6167311, SRR6167312 and SRR6167313 and under Gene Expression Omnibus accession number GSE109433.

References

29. Touraev, A. & Heberle-Bors, E. Microspore embryogenesis and in vitro pollen maturation in tobacco. *Methods Mol. Biol.* **111**, 281–291 (1999).
30. 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. USA* **107**, 16384–16389 (2010).
31. Carrier, G. et al. An efficient and rapid protocol for plant nuclear DNA preparation suitable for next generation sequencing methods. *Am. J. Bot.* **98**, e13–e15 (2011).
32. Vergne, P. et al. Somatic embryogenesis and transformation of the diploid rose *Rosa chinensis* cv 'Old Blush'. *Plant Cell Tissue Organ Cult.* **100**, 73–81 (2010).
33. Gnerre, S. et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. USA* **108**, 1513–1518 (2011).
34. 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).
35. Wang, C. et al. Genome-wide analysis of local chromatin packing in *Arabidopsis thaliana*. *Genome Res.* **25**, 246–256 (2015).
36. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
37. Akdemir, K. C. & Chin, L. HiCPlotter integrates genomic data with interaction matrices. *Genome Biol.* **16**, 198 (2015).
38. Tang, H. et al. ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol.* **16**, 3 (2015).
39. Chin, C. S. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
40. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
41. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
42. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
43. Foissac, S. et al. Genome annotation in plants and fungi: EuGene as a model platform. *Curr. Bioinform.* **3**, 87–97 (2008).
44. Schläpfer, P. et al. Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant Physiol.* **173**, 2041–2059 (2017).

© 2018 Nature America Inc., part of Springer Nature. All rights reserved.

45. Cottret, L. et al. MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks. *Nucleic Acids Res.* **38**, W132–W137 (2010).
46. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45 D1**, D158–D169 (2017).
47. Zheng, Y. et al. iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol. Plant* **9**, 1667–1670 (2016).
48. Dai, X., Sinharoy, S., Udvardi, M. & Zhao, P. X. PlantTFcat: an online plant transcription factor and transcriptional regulator categorization and analysis tool. *BMC Bioinformatics* **14**, 321 (2013).
49. Finn, R. D. et al. InterPro in 2017: beyond protein family and domain annotations. *Nucleic Acids Res.* **45 D1**, D190–D199 (2017).
50. Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering transposable element diversification in de novo annotation approaches. *PLoS One* **6**, e16526 (2011).
51. Hoede, C. et al. PASTEC: an automatic transposable element classification tool. *PLoS One* **9**, e91929 (2014).
52. Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
53. Quesneville, H. et al. Combined evidence annotation of transposable elements in genome sequences. *PLOS Comput. Biol.* **1**, 166–175 (2005).
54. Formey, D. et al. The small RNA diversity from *Medicago truncatula* roots under biotic interactions evidences the environmental plasticity of the miRNAome. *Genome Biol.* **15**, 457 (2014).
55. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42**, D68–D73 (2014).
56. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
57. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. Lond. Edinb. *Dublin Philos. Mag. J. Sci.* **2**, 559–572 (1901).
58. Jombart, T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405 (2008).
59. Salse, J. Ancestors of modern plant crops. *Curr. Opin. Plant Biol.* **30**, 134–142 (2016).
60. Adams, R. P. *Identification of Essential Oil Components By Gas Chromatography/Mass Spectrometry.*. 4th edn. (Allured Publishing Corporation, Carol Stream, IL, USA, 2007).
61. Veluchamy, A. et al. LHP1 regulates H3K27me3 spreading and shapes the three-dimensional conformation of the arabidopsis genome. *PLoS One* **11**, e0158936 (2016).
62. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
63. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
64. Zang, C. et al. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **25**, 1952–1958 (2009).
65. 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**, 576–589 (2010).
66. Ye, T. et al. seqMINER: an integrated ChIP-seq data interpretation platform. *Nucleic Acids Res.* **39**, e35 (2011).
67. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
68. 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).
69. 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).
70. Dubois, A. et al. Genomic approach to study floral development genes in Rosa sp. *PLoS One* **6**, e28455 (2011).
71. Marcial-Quino, J. et al. Stem-loop RT-qPCR as an efficient tool for the detection and quantification of small RNAs in *Giardia lamblia*. *Genes (Basel)* **7**, E131 (2016).

© 2018 Nature America Inc., part of Springer Nature. All rights reserved.

nature research

Corresponding author(s): BENDAHMANE Mohammed

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.

For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Experimental design

1. Sample size

Describe how sample size was determined.

No sample size calculation was required for this study.

2. Data exclusions

Describe any data exclusions.

No data exclusion

3. Replication

Describe the measures taken to verify the reproducibility of the experimental findings.

RNA-seq experiments were performed in triplicate. ChIP-seq experiments were performed on independent tissues. All real time quantitative RT-PCR were repeated at least 3 times.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

The study does not involve any randomized experimental group. Phylogenetic groups are based on published data.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

No blinding was required for this study.

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☒	☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☒	☐ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ A statement indicating how many times each experiment was replicated
☒	☐ The statistical test(s) used and whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☒	☐ Test values indicating whether an effect is present *Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.*
☒	☐ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☒	☐ Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	Code for filtering FALCON overlaps can be found at http://lipm-bioinfo.toulouse.inra.fr/download/til-r/. This is indicated in the manuscript.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.	The data availability statement is included in the Author Information section of the main manuscript

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	ChIP assays were performed using the commercially available antibodies, anti-H3K9ac (Millipore, ref. 07-352) and anti-H3K27me3 (Millipore, ref. 553 07-449), previously validated by other published data and widely used for similar experiments.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No cell lines were used in this study. The source of plants used in this study is given in Supplementary Table 7 (section 8 of Supplementary Notes).
b. Describe the method of cell line authentication used.	No cell lines were used in this study.
c. Report whether the cell lines were tested for mycoplasma contamination.	No cell lines were used in this study.
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No cell lines were used in this study.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.	No animals were used in this study.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	No human research was performed in this study.

nature research

Corresponding author(s): BENDAHMANE Mohammed

ChIP-seq Reporting Summary

Form fields will expand as needed. Please do not leave fields blank.

Data deposition

1. For all ChIP-seq data:

☒ a. Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ b. Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

2. Provide all relevant data deposition access links. *The entry may remain private before publication.*	ChipSeq raw data have been deposited under SRA numbers SRR6167310, SRR6167311, SRR6167312 and SRR6167313. Processed data can be accessed through genome browser https://lipm-browsers.toulouse.inra.fr/pub/RchiOBHm-V2/ (login: reviewer / password: rosaob).
3. Provide a list of all files available in the database submission.	RcHt_17_H3K9ac_mark_Petal.R1.fastq.gz RcHt_17_H3K9ac_mark_Petal.R2.fastq.gz RcHt_17_H3K9Ac_Input_Petal.R1.fastq.gz RcHt_17_H3K9Ac_Input_Petal.R2.fastq.gz RcHt_17_H3K27me3_mark_Petal.fastq.gz RcHt_17_H3K27me3_Input_Petal.fastq.gz
4. Provide a link to an anonymized genome browser session (e.g. UCSC), if available.	The genome and data can be accessible for the reviewer at https://lipm-browsers.toulouse.inra.fr/pub/RchiOBHm-V2/ (login: reviewer / password: rosaob).

Methodological details

5. Describe the experimental replicates.	n.a
6. Describe the sequencing depth for each experiment.	Homozygous genome : H3K9Ac mark, mean coverage : 5.2x,; H3K9Ac input, mean coverage : 15x; H3K27me3 mark, mean coverage : 4.9x; H3K27me3 input : 4.2x Heterozygous genome : H3K9Ac mark, mean coverage : 3.6x; H3K9Ac input, mean coverage : 10.2x; H3K27me3 mark, mean coverage : 3.4x; H3K27me3 input : 2.9x
7. Describe the antibodies used for the ChIP-seq experiments.	H3K9AC : MILLIPORE / 07-352 / n.a / 2325091 H3K27me3 : MILLIPORE / 07-449 / n.a / 2686928
8. Describe the peak calling parameters.	** H3K9AC mark : * read mapping : bowtie2 -x $INDEX_FILE -m1 $R1_file -m2 $R2_file -S $RESULT_FILE -N1 $INDEX_FILE = [RchiOBHm-V2.0,RcHzRDP12] - Chip : $R1_file = $DATA_REP/trim_galore/RcHt_17_H3K9ac_mark_Petal.R1.val1.fq.gz $R2_file = $DATA_REP/trim_galore/RcHt_17_H3K9ac_mark_Petal.R2.val2.fq.gz $RESULT_FILE = [aln_bowtie2_Rose_K9ac_mark_Flower_vs_RchiOBHm-V2.0.sam,aln_bowtie2_Rose_K9ac_mark_Flower_vs_RcHzRDP12.sam] - Input : $R1_file = $DATA_REP/trim_galore/RcHt_17_H3K9ac_Input_Petal.R1.val1.fq.gz $R2_file = $DATA_REP/trim_galore/RcHt_17_H3K9ac_Input_Petal.R2.val2.fq.gz $RESULT_FILE = [aln_bowtie2_Rose_K9ac_Input_Flower_vs_RchiOBHm-V2.0.sam,aln_bowtie2_Rose_K9ac_Input_Flower_vs_RcHzRDP12.sam] * Peak calling : macs2 callpeak -f BAM -t $mark_file -n $result_file -g 515588973 -c $control_file --bdg - RchiOBHm genome :

$mark_file = aln_bowtie2_Rose_K9ac_mark_Flower_vs_RchiOBHm-V2.0.bam
$result_file = aln_bowtie2_Rose_K9ac_Flower_vs_RchiOBHm-V2.0.macs2
$control_file = aln_bowtie2_Rose_K9ac_Input_Flower_vs_RchiOBHm-V2.0.bam
- RcHzRDP12 genome :
$mark_file = aln_bowtie2_Rose_K9ac_mark_Flower_vs_RcHzRDP12.bam
$result_file = aln_bowtie2_Rose_K9ac_Flower_vs_RcHzRDP12.macs2
$control_file = aln_bowtie2_Rose_K9ac_Input_Flower_vs_RcHzRDP12.bam
**H3K27me3 :
* read mapping :
bowtie2 -x $INDEX_FILE -U R_file -S $RESULT_FILE -N1
$INDEX_FILE = [RchiOBHm-V2.0,RcHzRDP12]
- Chip :
$R_file = RcHt_17_H3K27me3_mark_Petal.trimmed.fq.gz
$RESULT_FILE = [aln_bowtie2_Rose_H3K27me3_mark_Flower_vs_RchiOBHm-V2.0.sam,aln_bowtie2_Rose_H3K27me3_mark_Flower_vs_RcHzRDP12.sam]
- Input :
$R_file = RcHt_17_H3K27me3_Input_Petal.trimmed.fq.gz
$RESULT_FILE = [aln_bowtie2_Rose_H3K27me3_Input_Flower_vs_RchiOBHm-V2.0.sam,aln_bowtie2_Rose_H3K27me3_Input_Flower_vs_RcHzRDP12.sam]
* Peak calling :
macs2 callpeak -f BAM -t $mark_file -n $result_file -g 515588973 -c $control_file --bdg
- RchiOBHm genome :
$mark_file = aln_bowtie2_Rose_H3K27me3_mark_Flower_vs_RchiOBHm-V2.0.bam
$result_file = aln_bowtie2_Rose_H3K27me3_Flower_vs_RchiOBHm-V2.0.macs2
$control_file = aln_bowtie2_Rose_H3K27me3_Input_Flower_vs_RchiOBHm-V2.0.bam
- RcHzRDP12 genome :
$mark_file = aln_bowtie2_Rose_H3K27me3_mark_Flower_vs_RcHzRDP12.bam
$result_file = aln_bowtie2_Rose_H3K27me3_Flower_vs_RcHzRDP12.macs2
$control_file = aln_bowtie2_Rose_H3K27me3_Input_Flower_vs_RcHzRDP12.bam

9. Describe the methods used to ensure data quality.

** H3K9AC mark :
* RchiOBHm genome :
# peaks with FDR < 5% = 46624
# peaks FC > 5 = 21668
* RcHzRDP12 genome :
# peaks with FDR < 5% = 50577
# peaks FC > 5 = 27286
**H3K27me3 mark :
* RchiOBHm genome :
# peaks with FDR < 5% = 13877
# peaks FC > 2 = 6379
* RcHzRDP12 genome :
# peaks with FDR < 5% = 20179
# peaks FC > 5 = 11799

10. Describe the software used to collect and analyze the ChIP-seq data.

Software used to collect and analysed ChipSeq data are fully described in section 6.2 of Supplementary Notes

nature research

Corresponding author(s): BENDAHMANE Mohammed

Flow Cytometry Reporting Summary

Form fields will expand as needed. Please do not leave fields blank.

Data presentation

For all flow cytometry data, confirm that:

☒ 1. The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ 2. The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ 3. All plots are contour plots with outliers or pseudocolor plots.

☒ 4. A numerical value for number of cells or percentage (with statistics) is provided.

Methodological details

5. Describe the sample preparation.	Nuclei were isolated from RcHzRDP12 calli or from young leaves of regenerated plantlets, as previously described17, and stained by adding 1µg/ml DAPI (Sigma) for 1 hour at room temperature.
6. Identify the instrument used for data collection.	FACS data collections were performed using MACSQuant VYB (Miltenyi Biotec) cytometer.
7. Describe the software used to collect and analyze the flow cytometry data.	FACS data analyses were performed using FlowJo software (FlowJo LLC).
8. Describe the abundance of the relevant cell populations within post-sort fractions.	Total nuclei populations were gated using DAPI intensity (Extended Data Fig. 1j, left panels): 16% of 'Old blush' and 13% of RcHzRDP12 nuclei were retained as DAPI+. Next, doubles were eliminated from this population using DAPI-W versus DAPI-H axes (Extended Data Fig. 1j, middle panels): about 97% of DAPI+ nuclei were retained by this gating in both samples. Finally, in DAPI+ singles cells, the proportions of nuclei with different ploidy levels were determined based on their DAPI intensity (Extended Data Fig. 1j, right panels): about 67% and 12% of nuclei were diploids (2N) and tetraploids (4N), respectively in both the "Old Blush" sample and in the RcHzRDP12 sample.
9. Describe the gating strategy used.	Total nuclei populations were gated using DAPI intensity (Extended Data Fig. 1j, left panels). Next, doubles were eliminated from this population using DAPI-W versus DAPI-H axes (Extended Data Fig. 1j, middle panels). Finally, in DAPI+ singles cells, the proportions of nuclei with different ploidy levels were determined based on their DAPI intensity (Extended Data Fig. 1j, right panels).

Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information. ☒