OXFORD UNIVERSITY PRESS | Molecular Biology and Evolution

Duplication and specialization of *NUDX1* in *Rosaceae* led to geraniol production in rose petals

Journal:	*Molecular Biology and Evolution*
Manuscript ID	Draft
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Conart, Corentin; Université Jean Monnet Saint-Étienne Faculté des Sciences et Techniques, LBVpam Saclier, Nathanaelle; Universite Claude Bernard Lyon 1, Laboratoire d'Ecologie des Hydrosystèmes Naturels et Anthropisés Foucher, Fabrice; INRAE Angers, IRHS Goubert, Clement; McGill University, Rius-Bony, Aurélie; Université Jean Monnet Saint-Étienne Faculté des Sciences et Techniques, LBVpam Paramita, Saretta; Université Jean Monnet Saint-Étienne Faculté des Sciences et Techniques, LBVpam Moja, Sandrine; Université Jean Monnet Saint-Étienne Faculté des Sciences et Techniques, LBVpam Thouroude, Tatiana; INRAE Angers, IRHS Douady, Christophe; Universite de Lyon 1, CNRS UMR 5023, Laboratoire d'Ecologie des Hydrosystèmes Fluviaux Sun, Pulu; University of Amsterdam Swammerdam Institute for Life Sciences, Green Life Sciences Research Cluster Nairaud, Baptiste; Université Jean Monnet Saint-Étienne Faculté des Sciences et Techniques, LBVpam Saint-Marcoux, Denis; Université Jean Monnet Saint-Étienne Faculté des Sciences et Techniques, LBVpam Bahut, Muriel; INRAE Angers, IRHS Jeauffre, Julien; INRAE Angers, IRHS Hibrand Saint-Oyant, Laurence; INRAE Angers, IRHS Schuurink, Robert; University of Amsterdam Swammerdam Institute for Life Sciences, Green Life Sciences Research Cluster Magnard, Jean-Louis; Université Jean Monnet Saint-Étienne Faculté des Sciences et Techniques, LBVpam Boachon, Benoît; Université Jean Monnet Saint-Étienne Faculté des Sciences et Techniques, LBVpam Dudareva, Natalia; Purdue University, Landscape Horticulture Baudino, Sylvie; Université Jean Monnet Saint-Étienne Faculté des Sciences et Techniques, LBVpam Caissard, Jean-Claude; Université Jean Monnet Saint-Étienne Faculté des Sciences et Techniques, LBVpam
Key Words:	floral scent, terpenes, nudix hydrolase, Rosa

Article, Discoveries

Duplication and specialization of *NUDX1* in *Rosaceae* led to geraniol production in rose petals

Corentin CONART[1], Nathanaelle SACLIER[2], Fabrice FOUCHER[3], Clément GOUBERT[4], Aurélie RIUS-BONY[1], Saretta N. PARAMITA[1], Sandrine MOJA[1], Tatiana THOUROUDE[3], Christophe DOUADY[2,5], Pulu SUN[6], Baptiste NAIRAUD[1], Denis SAINT-MARCOUX[1], Muriel BAHUT[7], Julien JEAUFFRE[3], Laurence HIBRAND SAINT-OYANT[3], Robert C. SCHUURINK[6], Jean-Louis MAGNARD[1], Benoît BOACHON[1], Natalia DUDAREVA[8,9], Sylvie BAUDINO[1,*], Jean-Claude CAISSARD[1]

* Corresponding author: Sylvie.Baudino@univ-st-etienne.fr

[1]Université Lyon, Université Saint-Etienne, CNRS, UMR 5079, Laboratoire de Biotechnologies Végétales appliquées aux Plantes Aromatiques et Médicinales, F-42023 Saint-Etienne, France; [2]Université Lyon, Université Claude Bernard Lyon 1, CNRS, UMR 5023, ENTPE, Laboratoire d'Ecologie des Hydrosystèmes Naturels et Anthropisés, F-69622 Villeurbanne, France; [3]Univ Angers, Institut Agro, INRAE, IRHS, SFR QUASAV, F-49000 Angers, France; [4]Department of Human Genetics, McGill University Genome Center, 740 Dr Penfield Ave, Montreal, Quebec H3A 0G1, Canada; [5]Institut Universitaire de France, F-75005 Paris, France; [6]Green Life Sciences Research Cluster, Swammerdam Institute for Life Sciences, University of Amsterdam, Science Park 904, 1098 XH, Amsterdam, The Netherlands; [7]Univ Angers, SFR QUASAV, F-49000 Angers, France; [8]Department of Biochemistry, Purdue University, West Lafayette, IN 47907, USA; [9]Purdue Center for Plant Biology, Purdue University, West Lafayette, IN 47907, USA.

Keywords

Rosaceae, *Rosa*, Nudix hydrolase, monoterpenes, *NUDX1* synteny

Abstract

Nudix hydrolases are conserved enzymes ubiquitously present in all kingdoms of life. Recent research revealed that several Nudix hydrolases are involved in terpenoid metabolism in plants. In modern roses, RhNUDX1 is responsible for formation of geraniol, a major compound of rose scent. Nevertheless, this compound is produced by monoterpene synthases in many geraniol-producing plants. As a consequence, this raised the question about the origin of RhNUDX1 function and the *NUDX1* gene evolution in *Rosaceae*, in wild roses or/and during the domestication process. Here, we showed that three distinct clades of *NUDX1* emerged in the *Rosoidae* subfamily (Nudx1-1 to Nudx1-3 clades), and two subclades evolved in the *Rosa* genus (Nudx1-1a and Nudx1-1b subclades). We also showed that the Nudx1-1b subclade was more ancient than the Nudx1-1a subclade, and that the *NUDX1-1a* gene emerged by a *trans*-duplication of the more ancient *NUDX1-1b* gene. After the transposition, *NUDX1-1a* was *cis*-duplicated, leading to a gene dosage effect on the production of geraniol in different species. Furthermore, the *NUDX1-1a* appearance was accompanied by the evolution of its promoter, most likely from a *Copia* retrotransposon origin, leading to its petal-specific expression. Thus, our data strongly

suggest that the unique function of *NUDX1-1a* in geraniol formation was evolved naturally in the genus *Rosa* before domestication.

Introduction

Rosa is a complex taxon with more than 150 intertwined species (Wissemann 2003). Only few (around 15) rose species have been domesticated by humans since Antiquity (fig. 1). In Knossos (1700 B.C.), roses were painted with only few petals like wild briars (fig. 1a), while in Rome and Pompei (79 A.C.) they were presented with dozens of petals (fig. 1b), meaning that the domestication process had already started. Indeed, over the past three centuries, domestication resulted in flowers with hundreds of petals often with a strong fragrance. Some of the very ancient roses, approximately 1,000 to 2,000 years old, have come down to us as heritage roses (fig. 1c). This includes *Rosa chinensis* cv. 'Old Blush' (Old Blush) from China, which is likely a natural hybrid between wild species (Raymond et al. 2018). This rose has been largely used by breeders, and many modern roses probably have Old Blush as an ancestor. Other heritage roses have also been used for horticultural selection and hybridization with other varieties (supplementary table S1, Supplementary Material online). As a result, modern roses are an extended combination between alleles of different wild species, and alleles that appeared by spontaneous bud mutations.

One of the most important traits attracting humans to roses is their pleasant fragrance. Geraniol is one of the rose scent constituents, which contributes to the flower rosy note. In contrast to most plants, formation of this monoterpene in modern roses does not rely on a canonical biosynthetic pathway (Magnard et al. 2015) that involves a plastidial monoterpene synthase (Sun et al. 2016). Instead, a cytosolic Nudix hydrolase (RhNUDX1) converts geranyl diphosphate (GPP) to geranyl phosphate (GP), which in turn is dephosphorylated by uncharacterized phosphatase to geraniol.

Nudix hydrolases are conserved enzymes hydrolyzing nucleoside diphosphates linked to some moiety X. They are ubiquitously present in all kingdoms of life and were proposed to function as housecleaning enzymes involved in cell sanitation (McLennan 2013; Yoshimura and Shigeoka 2015; Srouji et al. 2017). However, recent research revealed that Nudix hydrolases can be involved in terpenoid metabolism in plants (Magnard et al. 2015; Henry et al. 2018; Li et al. 2020; Sun et al. 2020). Indeed, *Arabidopsis thaliana* Nudix hydrolase 1 (AtNUDX1) together with an isopentenyl kinase coordinately regulates the isopentenyl diphosphate (IPP) amount destined for higher-order terpenoid biosynthesis (Henry et al. 2015; Henry et al. 2018). Although AtNUDX1 is also able to efficiently dephosphorylate GPP and farnesyl diphosphate (FPP) *in vitro*, no geraniol nor (*E,E*)-farnesol was detected in this species (Chen et al. 2003). In contrast, RwNUDX1-2 from a cultivated hybrid of *R. wichurana* hydrolyzes specifically cytosolic FPP into farnesyl phosphate (FP) *en route* to (*E,E*)-farnesol formation (Sun et al. 2020). The fact that members of NUDX1 family could have diverse functions in different species raises the question about RhNUDX1 evolution, whether it is present only in cultivated modern roses, or was already evolved in wild *Rosa* and/or *Rosaceae* species.

Here, we investigated the origin of RhNUDX1 function. We analyzed the evolution of all *NUDX1* gene homologs, their genomic localization and synteny by comparing the recently published genomes of Old Blush (Hibrand Saint-Oyant et al. 2018; Raymond et al. 2018) and several closely related genomes in the *Rosaceae* family (fig. 1c). We also examined the transposable elements (TEs) surrounding these genes and proposed an evolutionary scenario of duplication and specialization of *NUDX1-1a*, the gene encoding the Nudix hydrolase responsible for the GPP hydrolysis in rose petals.

Results

RcNUDX1 is present in multiple copies in Old Blush, but only *RcNUDX1-1a* is highly expressed in its petals.

Discovery of terpene synthase-independent pathway for geraniol biosynthesis in modern roses and the involvement of *RhNUDX1* in its formation (Magnard et al. 2015) raised the question of how this trait was evolved. Thus, we have isolated the corresponding genomic sequence from *R.* x *hybrida* cv. 'Papa Meilland', which revealed that *RhNUDX1* contains a single intron (*RhNUDX1-rs* for reference sequence). This sequence was used for phylogenetic analysis of *NUDX1* genes in *Rosaceae* family. A Maximum Likelihood tree (ML tree) rooted with the *A. thaliana* homolog, *AtNUDX1,* was constructed using genomic sequences of Old Blush, *Fragaria vesca*, *Malus* x *domestica,* and *Prunus persica*, available in the Genome Database for *Rosaceae* (GDR, www.rosaceae.org, (Jung et al. 2019); supplementary table S2, Supplementary Material online) as well as recently published *R.* x *wichurana* sequences (Sun et al. 2020) (fig. 2). For the readability of the ML tree, we did not use Old Blush sequences that were 100% identical between them (supplementary table S3, Supplementary Material online).

This ML tree revealed three well-resolved at nearly all node clades, numbered Nudx1-1 to Nudx1-3, and a lesser-supported clade named Nudx1-4. Two sequences (*Prupe.1G302800* and *MD13G1049100*) could not be assigned to a clade, and appeared on branches with low bootstraps. Interestingly, these branches and the Nudx1-4 clade include exclusively sequences of *M.* x *domestica* and *P. persica*, while the three other clades contain all the sequences of Old Blush, *R.* x *wichurana* and *F. vesca*. As *M.* x *domestica* and *P. persica* belong to *Amygdaloidae* subfamily and *Rosa* species and *F. vesca* belong to *Rosoideae* subfamily (Xiang et al. 2017) (fig. 1c), it suggests that duplications of the first ancestral *NUDX1* ortholog led to divergent sequences in the Nudx1-4 clade in *Amygdaloidae*, but to homologous sequences in well-supported Nudx1-1 to Nudx1-3 clades in *Rosoideae. RhNUDX1-rs*, which is involved in geraniol production in horticultural roses, was found in the Nudx1-1 clade. This clade also encompasses closely related *RcNUDX1-1* sequences from Old Blush with 97.1 to 97.6% identity to the reference *RhNUDX1-rs* (supplementary table S2, Supplementary Material online), indicating that they could be the result of very recent duplications of the same gene.

To gain insights in the evolution of these paralogs, we analyzed their genomic organization in three Old Blush genomes published in the GDR (supplementary table S2, Supplementary Material online). We also sequenced the Old Blush accession using MinION technology (supplementary table S5, Supplementary Material online). This technology increases the error rate in sequences, but allows to

obtain very long reads without informatics assembly (Lu et al. 2016), thus to verify gene clusters on chromosomes 2 and 4, and also to detects alleles and null alleles on homologous chromosomes. Comparison of all these sequences allowed to draw a comprehensive map in Old Blush (fig. 3), and a synteny map in *Rosaceae* (fig. 4). Two clusters containing *NUDX1* paralogs were found in Old Blush genome. The first cluster on chromosome 4 included the more ancient gene, *RcNUDX1-3*, along with one copy of both *RcNUDX1-1b* and *RcNUDX1-2a,* but a pseudogene $^{\Psi}$*RcNUDX1-2a* with two STOP codons on the other homologous chromosome 4. The second cluster was on chromosome 2 and contained four nearly identical copies of *RcNUDX1-1a* and one pseudogene $^{\Psi}$*RcNUDX1-1a* with a STOP codon. The four copies are nearly identical showing 98.7% of DNA identities and 96.8 to 99.0% of protein identities (supplemental tables S3 and S4, Supplementary Material online). Surprisingly, the *RcNUDX1-1a* genes were totally absent on a second homologous chromosome 2, which thus correspond to a null allele ^{na}*RcNUDX1-1a*. Copies of *NUDX1-2* (*RcNUDX1-2b* and *RcNUDX1-2c*) were also found on chromosome 6 and 7, respectively.

Comparisons of the two *NUDX1* clusters and the surrounding genes of the other *Rosaceae* (fig. 4, and supplementary table S2, Supplementary Material online) revealed that the possible ancestral gene *NUDX1-3* has duplicated on chromosome 4 thus separating *Amygdaloidae* and *Rosoideae* subfamilies, and giving respectively sequences of the Nudx1-4 clade, and Nudx1-1 and Nudx1-2 clades. Indeed, they were in the same microsyntenic region (fig. 4a). Furthermore, the two unresolved sequences *Prupe.1G302800* and *MD13G1049100* (fig. 2) were close to the homolog of the marker gene F, in similar position to *RcNUDX1-3* and its orthologs in *F. vesca* and *P. micrantha*, implying that the ancestral gene had highly diverged between *Rosoideae* and *Amygdaloidae*. The other cluster, with the *RcNUDX1-1a* copies, was unique to Old Blush, indicating that it likely had evolved in very ancient roses at the beginning of the domestication process or in wild ancestors of Old Blush (fig. 4b).

Our previous RNAseq, QTL and correlation analyses (Magnard et al. 2015; Sun et al. 2020) performed mainly on modern hybrid roses, showed that *RcNUDX1-1a* was expressed in petals and responsible for the geraniol production. On the other hand, the RcNUDX1-1b protein was active *in vitro*, but the *RcNUDX1-1b* gene was not expressed. We verified that it was also the case in a wild species by checking the *in vitro* activities of RmNUDX1-1a and RmNUDX1-1b proteins of the Moschata accession. These activities were quite similar to those of the corresponding Old Blush enzymes (supplementary table S6, Supplementary Material online), suggesting that only the gene expression could be responsible of geraniol production in wild species. Thus, to determine whether the other *RcNUDX1-1a* homologs*, RcNUDX1-1b, RcNUDX1-2 and RcNUDX1-3*, were expressed in petal tissue, qRT-PCR analyses with gene-specific primers were performed (supplementary table S7, Supplementary Material online). These analyses revealed that only *RcNUDX1-1a* transcripts indeed accumulate at high levels in Old Blush petals (60,000x more than *RcNUDX1-1b*), thus further suggesting that such mode of expression is rose specific and uniquely clustered *RcNUDX1-1a* paralogs are involved in the biosynthesis of geraniol (supplementary fig. S1, Supplementary Material online).

Taken together, these results support that the *NUDX1-3* ancestral orthologs were duplicated many times in the *Rosaceae*. The ortholog was probably an ortholog of *AtNUDX1* that had likely the same function. While genes within the Nudx1-1 and Nudx1-2 clades evolved in the subfamily *Rosoideae*, the *NUDX1-1a* paralogs emerged only in the genus *Rosa*. In addition, the high sequence similarity of the clustered *RcNUDX1-1a* paralogs with the characterized *RhNUDX1-rs,* as well as high level of expression, suggest that these paralogs are involved in the biosynthesis of geraniol in Old Blush. The presence of ^{na}*RcNUDX1-1a* opens the possibility that one of the potential wild parents of Old Blush did not have such cluster, and therefore the duplication of *RcNUDX1-1a* had occurred in wild species of the genus *Rosa*.

The *NUDX1-1a* paralogs are specific to wild roses producing geraniol.

To determine whether *RcNUDX1-1a* had already arisen in wild species of *Rosa* or evolved early during the domestication process, we performed GC-MS metabolic profiling of the volatiles produced by petals along with analysis of the *RcNUDX1-1* homologs in a collection of 29 accessions of wild roses and six accessions of heritage roses (supplementary tables S1, and S8, Supplementary Material online). Their genomic DNAs and mRNAs were used to isolate and characterize full-length *NUDX1-1* sequences (table 1, and supplementary table S9, Supplementary Material online). Due to the high sequence identity (89.5 to 91.6%, supplementary table S3, Supplementary Material online) between *RcNUDX1-1a* and *RcNUDX1-1b*, the primers were designed based on Old Blush sequences to amplify the region from ATG to STOP codons (supplementary table S7, Supplementary Material online). Sequencing of the obtained PCR products revealed that the primers were specific for Nudx1-1 clade and did not amplify sequences of the Nudx1-2 and Nudx1-3 clades.

cDNAs were obtained from all species that emit geraniol except for *R. sericea* producing a very small amount of this compound (supplementary tables S8, and S9, Supplementary Material online). We also cloned cDNAs from *R. rubus* that does not produce geraniol, but these cDNAs were as close to *RcNUDX1-1a* as to *RcNUDX1-1b*. For most of the accessions, several genomic sequences (gDNA) of *NUDX1-1* were obtained. However, numerous gDNAs were attained for some species due to the ploidy level (table 1, see supplementary table S1 for ploidy levels, Supplementary Material online) and two species have only a single gDNA. Interestingly, in *R. rubus*, no *NUDX1-1* gDNA corresponding to the isolated cDNAs was detected.

All identified gDNAs contained one intron of variable size (supplementary table S2, Supplementary Material online), and clustered in two groups on the ML tree (fig. 5, and supplementary fig. S2, Supplementary Material online). The first group included the Old Blush *RcNUDX1-1a*, and thus was named Nudx1-1a subclade (orange names on supplementary fig. S2, Supplementary Material online), while the second group, named Nudx1-1b subclade, included the gDNAs which were closer to *RcNUDX1-1b* than to *RcNUDX1-1a* (red names on supplementary fig. S2, Supplementary Material online). A blastn analysis of all the gDNAs (supplementary table S9, Supplementary Material online) revealed that most of the gDNAs on the ML tree share 88.7% to 99.8% identity with both the *RcNUDX1-1a* and *RcNUDX1-1b* sequences. Thus, gDNAs displaying identity more than 1% higher with *RcNUDX1-1a* than with *RcNUDX1-1b* were assigned to the Nudx1-1a subclade and *vice versa* (supplementary fig.

S2, Supplementary Material online). Few gDNAs were as close to *RcNUDX1-1a* as to *RcNUDX1-1b* since they exhibit less than 1% identity in favor to either of two subclades (shown in black on supplementary fig. S2, Supplementary Material online). These sequences were often distant from all other gDNAs (long black branches on supplementary fig. S2, Supplementary Material online) and could have thus diverged in these particular species. Some of them were located at the root of the tree suggesting that they could represent *NUDX1-1* ancestral sequences.

In contrast to Nudx1-1a subclade, Nudx1-1b subclade included all the gDNAs from the species that don't produce geraniol (blue stars in fig. 5, table 1, and supplementary table S8, Supplementary Material online). Unlike *NUDX1-1a* gDNAs, which were clearly absent in 8 accessions, *NUDX1-1b* gDNAs were undetectable only in 2 accessions (supplementary table S9, Supplementary Material online). The gDNAs of Nudx1-1b subclade were closer to the root of the phylogenetic tree than those of the Nudx1-1a subclade. Thus, despite weak branch support of the ML tree, these data suggest an ancestral origin of the *NUDX1-1b* genes.

All cloned cDNAs were found to correspond to the ORF sequence found only in gDNAs belonging to the Nudx1-1a subclade (orange asterisks on fig. 5, supplementary table S9, Supplementary Material online), suggesting that only members of this clade are expressed. Next, we evaluated expression of *NUDX1-1* homologs in the petals of all 34 accessions (table 1, and supplementary tables S1, and S10, Supplementary Material online) by qRT-PCR with consensus primers, which were capable of amplifying both *NUDX1-1a* and *NUDX1-1b* (supplementary table S7, Supplementary Material online). As no cDNAs belonging to the *NUDX1-1b* group were obtained, transcripts detected in this analysis correspond to *NUDX1-1a* homologs (table 1). *NUDX1-1* transcripts were barely detected in botanical species not producing geraniol. In contrast, *NUDX1-1* was expressed in all species producing geraniol and for which genomic sequences corresponding to *NUDX1-1a* were obtained. The exceptions include two geraniol-producing species (accessions Hugonis B and Ecae) with very low *NUDX1-1* expression, and two low geraniol producers (accessions Foetida and Persian Yellow) with substantial *NUDX1-1* expression (table 1). In the latter two species, low geraniol levels could be the result of substrate limitation, while in two former species another *NUDX1* homolog could be involved in geraniol production. We have recently shown the existence of specialization of different homologs as *RwNUDX1-2c* was active in *R.* x *wichurana*, but not in Old Blush (Sun et al. 2020). In botanical and heritage roses, *NUDX1-1a* expression was highly correlated (*P*-values < 0.001) with geraniol levels, as well as with the levels of acyclic monoterpenes (supplementary fig. S3, and supplementary table S11, Supplementary Material online). It was also positively correlated with the production of the acyclic sesquiterpenes (*E,E*)-farnesol, (*E,E*)-α-farnesene and (*Z,E*)-α-farnesene as well as 2-phenylethanol. A negative correlation was found for 2-pentadecanone.

Thus, the presence of *NUDX1-1a* paralogs and its expression in some but not all botanical species as well as a positive correlation between *NUDX1-1a* expression and geraniol levels could indicate that the unique function of *NUDX1-1a* in geraniol production was evolved naturally in the genus *Rosa* before domestication.

Trans-duplication of _NUDX1-1b_ and additional _cis_-duplications led to a _NUDX1-1a_ cluster in the genus _Rosa_.

Our data show that the ancestral *RcNUDX1-1b* gene homologs exist in many wild roses and in some other *Rosaceae* species, while *RcNUDX1-1a* homologs are only present in some wild roses mostly producing geraniol. This strongly suggested that *NUDX1-1a* homologs arose from *trans*-duplication of *NUDX1-1b* in wild roses, followed by *cis*-duplications on chromosome 2.

To understand the origin of the clustered *RcNUDX1-1a* paralogs on chromosome 2, we first performed a dot-plot analysis of nucleotide sequence similarity (supplementary fig. S4, Supplementary Material online). The identified repeated sequences (supplementary fig. S4a, Supplementary Material online) were then compared to the TEs annotated in the GDR (supplementary fig. S4b, and supplementary table S12, Supplementary Material online) to draw a comprehensive map (fig.6). This analysis revealed that all five copies of *RcNUDX1-1a* with their intergenic regions were nearly identical and contained the same TEs in the same order (fig. 6a). Each *NUDX1-1a* copy was surrounded by a fragment of the *Copia R24588* retrotransposon (Class I, RNA intermediate) at the 5'-end, and by two embedded Miniature Interspersed TEs (MITEs; Wicker et al. 2007) at the 3'-end (except for copy #5). MITE *G13554* itself was inserted into MITE *P580.2030* (respectively named in the GDR as *ms382250_RcHm_v2.0_Chr2_DXX-MITE_denovoRcHm_v2.0-B-G13554-Map6* and *ms580616_RcHm_v2.0_Chr2_noCat_denovoRcHm_v2.0-B-P580.2030-Map20*). The embedded MITEs in the second copy were interrupted by a long sequence containing genes, non-coding RNAs, and TEs (supplementary table S12, Supplementary Material online). Analysis of the four copies of these embedded MITEs revealed that they all have more than 80% of identity compared to their consensus sequences published in the GDR (supplementary table S12, Supplementary Material online), suggesting that the initial *RcNUDX1-1a* block may have then been duplicated in tandem after its initial insertion on chromosome 2.

To further analyze the origin of these block duplications, we searched for MITE *G13554,* MITE *P580.2030,* and *Copia R24588* localizations around the *RcNUDX1* homologs on other chromosomes, and found two copies on chromosome 4 (supplementary table S12, Supplementary Material online). Analysis of available genomic sequences of the two rose haplotypes of the GDR revealed that *Copia R24588* was absent on chromosome 4 of one annotated haplotype (Raymond et al. 2018), while it was found manually in the other (Hibrand Saint-Oyant et al. 2018). To compare the organization of the clusters on chromosomes 2 and 4 in different species, we also performed MinION sequencing of Moschata accession, which produces geraniol, and of Laevigata accession, an unscented rose species (supplementary table S13, Supplementary Material online). In Moschata, we found two copies of *RmNUDX1-1a* harboring the same organization of TEs as in Old Blush, but none in the accession Laevigata (fig. 6a). As *R. laevigata* is more ancient than *R. moschata*, which in turn is more ancient than *R. chinensis* cv. 'Old Blush' (Fougère-Danezan et al. 2015; Debray et al. 2019), these results suggest that a series of duplications occurred during the evolution of the genus *Rosa*. Analysis of microsyntenic region of chromosome 4, that includes the cluster *RcNUDX1-3/RcNUDX1-1b/RcNUDX1-2a,* revealed a

sequence *NUDX1-1b* directly upstream of the same MITE and *Copia R24588* elements found in the chromosome 2 of Old Blush and Moschata (fig. 6b). Contrary to chromosome 2, the MITE *P580.2030* was repeated in tandem and did not embed MITE *G13554*. The absence of the embedded MITE suggests that the *NUDX1* cluster on the chromosome 4 of Old Blush is a likely candidate for being the ancestral sequence from which *RcNUDX1-1a* blocks on chromosome 2 originate.

To determine whether in general *Rosa* species have multiple copies of *NUDX1-1a*, we estimated the copy number of *NUDX1-1* homologs in some wild roses using qPCR experiments on genomic DNA (Axelsson et al. 2013) (supplementary table S7, Supplementary Material online). Quantification was done for 12 wild species, and revealed that the number of *NUDX1-1a* copies ranged from three to ten in geraniol producing species and from two to five in species producing no geraniol (supplementary fig. S5, Supplementary Material online). These results clearly show that the number of *NUDX1-1* copies is indeed variable in rose species and overall higher in species producing geraniol.

Taken together, these results are consistent with a *trans*-duplication occurring in the genus *Rosa* between chromosome 4 and chromosome 2, and show that *NUDX1-1a* was a result of specialized duplication of *NUDX1-1b*. After this duplication, MITE *G13554* was inserted into MITE *P580.2030*. The sequence block *Copia R24588 NUDX1-1a* with MITE *P580.2030* [MITE *G13554*] at the beginning or at the end, was further duplicated in tandem in some wild roses producing geraniol.

Promoter specificity and gene dosage determine the high *NUDX1-1a* expression level in petals.

Our results indicate that the clustered *NUDX1-1a* paralogs arose from the duplication of the *NUDX1-1b* gene, which does not express in petals, raising the question of how tissue specificity and high levels of *NUDX1-1a* expression were achieved.

To answer this question, we first tested our hypothesis that a gene dosage affects *NUDX1-1a* expression in wild roses producing geraniol. Thus, we analyzed whether the number of *NUDX1-1* copies in the 13 already analyzed wild species (supplementary fig. S5, Supplementary Material online) correlates with the expression levels of *NUDX1-1* homologs (table 1). Indeed, the *NUDX1-1a* copy number positively correlated, although not linearly, with the expression of *NUDX1-1a* in rose petals (fig. 7). These results suggest that the number of duplication events leading to multiple copies of *NUDX1-1a* paralogs directly impacts its expression in petals. We did not try to find the exact expression level of each of the four copies of *RcNUDX1-1a*, because of the very high DNA sequence identities in the exons (Align_OldBlush_DNAsequences.fasta, and Clones_IntronExonStructure.fasta, Supplementary Material online), which would make almost impossible qRT-PCR experiment, even with a High Melting Resolution technique (Roccia et al. 2019). It was also because of the same length and structure of their promoters (see below, and supplemental fig. S6, Supplementary Material online) which could indicate a similar expression.

Next, to investigate the contribution of promoters to different expression levels of the *RcNUDX1-1a* and *b* paralogs, we searched for the presence of specific sequences or structures upstream the coding sequences. In Old Blush, we manually identified four repeats of a conserved 38 bp sequence, designated as *box38* A to D. These repeats were identical in all five blocks of *RcNUDX1-1a*, #1 to #5,

and always located 138 bp upstream the *RcNUDX1-1a* transcription starting site. Moreover, we found a 33 bp overlap between *box38* A and a fragment of the *Copia R24588* localized at the at the 5'-end of each *NUDX1-1a* copy. In order to test the relationship between *Copia R24588* and *box38*, the fragments of *Copia R24588* and the *box38* repeats upstream of each copy of *RcNUDX1-1a* gene on the chromosome 2 were analyzed. The *Copia R24588* fragments contained the consensus sequence published in the GDR and identified from the interspersed copies of *Copia R24588* in the Old Blush genome. A search for short homologous sequences of *box38* in the Old Blush genome using blatsn and multiple sequence alignment (supplemental fig. S6, Supplemental Material online) confirmed that *box38* was the result of the 3'-end duplication of the *Copia R24588* fragment (supplemental fig. S6a, Supplemental Material online). There were no other *box38* elements in the Old Blush genome, but only very short fragments were found in other TE, intron, and intergenic hits (supplemental fig. S6b, Supplemental Material online). The available online PlantCARE tool (Lescot et al. 2002), was unable to detect any known binding sites for transcription factors in the *box38* repeats, which does not exclude the existence of unknown ones. To go further, we performed another multiple sequence alignment using the *Copia R24588* consensus sequence of the GDR. On this sequence, we aligned the following sequences: the *Copia R24588* fragment upstream *RcNUDX1-1a* blocks on chromosome 2, and the *Copia R24588* fragment upstream $^{\Psi}$*RcNUDX1-2a* on chromosome 4 (Fig. 8). The alignment clearly showed the origin of the promoter fragment (fig. 8a) in the complete consensus map of *Copia R24588*, with *box38* A being the best aligned within the 3' Long Terminal Repeat (LTR) of *Copia R24588* (fig. 8b). It also showed that *box38* B to D only exist upstream *RcNUDX1-1a* blocks (fig. 8c).

To find whether this pattern is conserved in botanical roses and important for the expression of *NUDX1-1a* in petals, we compared the upstream sequences of *NUDX1-1a* and *b* in a set of botanical roses producing and not producing geraniol (supplementary fig. S7, Supplementary Material online). While the number of *box38* repeats varied in the wild roses, the 138 pb distance between the last *box38* sequence and the ATG codon of the *NUDX1-1a* was conserved (supplementary fig. S7a, Supplementary Material online). In contrast, none of the upstream region of *NUDX1-1b* contained any *Copia R24588* sequence or *box38* repeats (supplementary fig. S7b, Supplementary Material online). One copy of the *box38* was also present in the *Copia R24588* elements upstream $^{\Psi}$*RcNUDX1-2a,* $^{\Psi}$*RmNUDX1-2a* and $^{\Psi}$*RlNUDX1-2a* pseudogenes on chromosome 4 suggesting that it could be more ancestral than those of chromosome 2.

All these results suggested a chronology of duplications: the *Copia R24588* fragment of chromosome 4 was *trans*-duplicated on chromosome 2, the *box38 A* was then *cis*-duplicated into four copies, and one of the putative blocks of fig. 6a was *cis*-duplicated on chromosome 2. Furthermore, these results indicated that the promoter of *RcNUDX1-1a* seemed to be unique, and originated from a specialization of a fragment of the LTR of *Copia R24588*.

Finally, we analyzed the impact of the *box38* repeats and different TEs in the promoter region of *RcNUDX1-1a* on the specific expression of this paralog in rose petals (fig. 9). Reporter gene encoding the green fluorescence protein (GFP) was fused to the promoter region of *RcNUDX1-1a* of different

lengths (fig. 9a). The longest *RcNUDX1-1a* promoter construct (*a1085:GFP*) included the entire 5'-region between MITEs and *RcNUDX1-1a* copy #4. The other constructs were made by removing the TEs one by one by PCR (supplementary table S7, Supplementary Material online). The *35S:GFP* used as a positive control displayed GFP fluorescence in parenchymous and epidermal cells (fig. 9b,and c). No detectable *GFP* expression was found in rose petals transferred with the empty vector (fig. 9d, and e) and the *RcNUDX1-1b* construct (1529 pb upstream of the ATG codon, named *b1529:GFP* construct) used as a negative control (fig. 9f,and g). GFP fluorescence was observed in rose petals expressing the three *RcNUDX1-1a* constructs, *a1085:GFP*, *a521:GFP*, and *a316:GFP* (fig. 9h to j). However, the removal of the *box38* repeats in the *a138:GFP* construct eliminated GFP expression (fig. 9k,and l) suggesting that the *box38* repeats are essential for petal expression.

Overall, these data suggest that the appearance of the *NUDX1-1a* paralogs by the transposition of *NUDX1-1b* was accompanied by the evolution of its promoter, likely by duplication of sequence in the LTR region of *Copia R24588*, leading to the specific expression of this paralogs in petals. This could come from the promoter of an ancestral copy of *NUDX1-2* which already added the *box38* fragment.

Discussion

Our analysis of the *NUDX1* genes in the *Rosaceae* family revealed that three clades (Nudx1-1 to Nudx1-3) evolved in the *Rosoidae* subfamily (including *P. micrantha*, *F. vesca* and *Rosa* species), and that two subclades (Nudx1-1a and Nudx1-1b) evolved in the *Rosa* genus (figs. 1, 2, 5, S2, and supplementary table S2, Supplementary Material online). Considering *AtNUDX1* as an outgroup and *RhNUDX1-rs* from a modern garden rose, the Nudx1-3 clade appeared to be more ancient than the others, and the Nudx1-1a subclade more recent. Comparative analysis of genetic maps of Old Blush, as a heritage rose producing geraniol, Moschata, as an accession of a wild rose producing geraniol, and Laevigata, as an accession of an unscented wild rose, allowed to access a global history of duplications in the *Rosoidae* subfamily (figs. 3, 4, and 6). The cluster *NUDX1-3/NUDX1-1b/NUDX1-2a* on chromosome 4 was found in *Rosoidae* accessions, suggesting a very old duplication of the putative ancestral *NUDX1-3* gene. In the *Amygdaloidae* subfamily (including *P. persica* and *M.* x *domestica*), their multiple copies in the same microsyntenic region (between marker genes F and Q on fig. 4) have significantly diverged, thus forming a different clade, Nudx1-4 (fig. 2). In contrast, the cluster of *NUDX1-1a* copies on chromosome 2 is more recent, specific to some species of the *Rosa* genus and absent in ancestral species like *R. banksiae*, *R. roxburghii* and *R. laevigata* (fig. 1c, and supplementary fig. S2, Supplementary Material online) (Fougère-Danezan et al. 2015; Debray et al. 2019). Moreover, the number of *NUDX1-1a* copies varies depending on species, with two copies in the Moschata accession, and five copies in Old Blush, for example (fig. 6, and supplementary fig. S5, Supplementary Material online). In Old Blush we identified two alleles on chromosome 2, one with five copies of *RcNUDX1-1a,* and the other with a null allele (fig. 3, and supplementary table S2, Supplementary Material online), which could confirm the previously predicted hybrid origin of this heritage rose (Raymond et al. 2018).

Our analysis of the TE landscape of *NUDX1-1* genes suggested a *trans*-duplication of a first paralog from chromosome 4 to 2, and then several *cis*-duplications of *NUDX1-1a* blocks including TEs in tandem (figs. 6, 10, supplementary fig. S4, and supplementary table S12, Supplementary Material online). The presence of TEs in both the putative source of *NUDX1-1a* on chromosome 4 and duplication blocks on chromosome 2 raise the possibility of a TE-mediated mechanisms. Indeed, sequence similarity between TE copies across the genome can be responsible for non-homologous recombination and the relocation and rearrangement of genomic features between TE dense regions (Cerbin and Jiang 2018), as observed for other biosynthetic gene clusters in plants (Boutanaev and Osbourn 2018). Further extensive analysis of the repeat content in *Rosa* species and other *Rosaceae* will be required to test this hypothesis and other putative TE-derived mechanisms, such as Pack-MULE or retrotransposition for example (Jiang et al. 2004; Cerbin and Jiang 2018; Krasileva 2019).

RcNUDX1-1a copies 2, 3 and 4 were found on chromosome 2 as repeats of a sequence block *Copia R24588 / RcNUDX1-1a* with MITE *P580.2030* [MITE *G13554*] at the beginning or at the end (fig. 6, and supplementary fig. S4, Supplementary Material online). In addition, *MITE P580.2030*, *Copia R24588*, and *NUDX1* homologs were found on one homologous chromosome 4 in a different configuration (*RcNUDX1-1b* / *MITE P580.2030* / *MITE P580.2030* /…/ *Copia R24588* / $^{\Psi}$*RcNUDX1-2a*) where *MITE P580.2030* does not include *MITE G13554,* but is *cis*-duplicated in tandem. This suggests that the copies of *NUDX1* on chromosome 4, including uninterrupted *MITE P580.2030*, are ancestral to those on the chromosome 2 and have been rearranged upon duplication (fig. 6). The parental status of the sequences on chromosome 4 is also supported by the fact that the microsynteny was not shared between *Rosa* species on chromosome 2 (five interspersed copies of *RcNUDX1-1a* in Old Blush, two in Moschata, and none in Laevigata), but was conserved on chromosome 4. Finally, high expression of *NUDX1-1a,* but *NUDX1-1b*, in petals of fully-opened flowers (table 1, supplementary fig. S1, and supplementary table S10, Supplementary Material online), further indicates that the cluster on chromosome 2 acquired petal-specific expression following duplication from chromosome 4 and subsequent duplication in tandem of the rearranged block. Such *cis*-duplications can occur by non-allelic homologous recombination between two identical sequences that may create an unequal crossing-over, or by microhomology-mediated break-induced replication mechanisms (Żmieńko et al. 2014; Lye and Purugganan 2019), even in synergy with TE mechanisms of translocation (Krasileva 2019). In *M.* x *domestica*, clusters of *O-METHYLTRANSFERASE* genes are associated with hairpins structures from palindromic TEs provoked by DNA slippage during replication (Han et al. 2007). In our work, *MITEs P580.2030* and *G13554* are also forming ~300-400 bp palindromes associated with each replicated *RcNUDX1-1a* block on chromosome 2.

We also discovered that repeats of a 38 bp fragment derived from the LTR region of *Copia R24588,* and named *box38*, was necessary and sufficient to drive previously discovered petal-specific *NUDX1-1a* expression in petals of fully opened flowers (Magnard et al. 2015) (figs. 7, 9, and supplementary fig. S1, Supplementary Material online). The *Copia R24588* / *box38* location in the 5'-upstream regions of the pseudogenes $^{\Psi}$*NUDX1-2a* suggests that this gene may have been expressed originally. Thus, even if it

really looks like a neofunctionalization process, one cannot exclude subneofunctionalization as well (see review in Baudino et al. 2020). However, during *trans*-duplication from chromosome 4 to chromosome 2, the *box38* repeats were shuffled and ended up in front of *NUDX1-1a* making its expression petal-specific (fig. 10). To date, there is increasing evidence that TEs are a source of diversification of species and can modify gene expression, particularly in the *Rosaceae* (Gu et al. 2016; Wang et al. 2016; Zhao et al. 2016; Daccord et al. 2017; Jiang et al. 2019). Examples include recurrent blooming of roses and strawberries due to an insertion of another *Copia* element in the intron 2 of the anti-florigen homolog *KSN* (Iwata et al. 2012), and formation more than five petals in roses due to insertion of an uncharacterized TE in the intron 8 of *APETALA2/TOE*, which deregulated its expression (Hibrand Saint-Oyant et al. 2018). Several TE insertions in promoters have also been described in *Rosaceae*, which modified transcription levels as a result of new binding sites for transcription factors or disruption of existing ones, new methylation/acetylation pattern, or hairpin structure formation (Han and Korban 2007; Wang et al. 2009; Gu et al. 2016; Morata et al. 2018; Ono et al. 2018; Zhang et al. 2019).

Our results show that *box38* is part of the LTR region of *Copia R24588*. LTR flank the internal coding region of LTR retrotransposons and act as promoter for the selfish transcription of the canonical elements of the retrotransposon. LTR regions contain regulatory sequences that can modify gene expression occurring in *cis* and can contribute to neofunctionalization in plants and eukaryotes (Kobayashi et al, 2004, Grandbastien 2015, Galindo-González et al. 2017). As Old Blush is rich in TEs, which constitute 63.2% of the genome including 35.2% of class I LTR retrotransposons (Hibrand Saint-Oyant et al. 2018), further investigations are necessary to understand the underlying mechanisms of petal-specific expression.

We also found that the number of *NUDX1-1a* copies impacts the level of geraniol emission in wild roses, in a non-linear gene dosage effect (fig. 7, and supplementary fig. S3, Supplementary Material online). A similar situation was described in mammals, where the copy number of genes encoding amylase was higher in populations with high-starch diets, but not strictly linearly correlated to the amylase concentration in saliva (Perry et al. 2007; Axelsson et al. 2013). In an evolution perspective, if the number of copies increases fitness, these copies can be fixed by adaptive natural selection rather than diverged by genetic drift (Hahn 2009). As *RcNUDX1-1a* copies are very similar to each other (96.8 to 99.0% of DNA identity resulting in 98.7% of protein identity), it is possible that this gene, and thus geraniol concentration, were important in the adaptation and evolution of *Rosa* species. Interestingly, the blocks on chromosome 2 in *Rosa* look similar to the repetitions of *MATE1* in *Zea mays,* which include copies of *Copia*, *Gypsy*, and *Mutator* in their intergenic regions and for which the total number of gene copies is associated with aluminium tolerance (Maron et al. 2013). This polymorphism is referred as Copy Number Variations (CNVs), i.e. variation of number of gene copies between individuals (Lye and Purugganan 2019), or between inbred lines (Maron et al. 2013). It has been demonstrated that such CNVs could be a very strong driving force leading to adaptations (DeBolt 2010) even *via* secondary metabolism (Prunier et al. 2017; Shirai and Hanada 2019). The differences of copy number between Old Blush, Moschata, and Laevigata (figs. 6, 7, and supplementary fig. S5, Supplementary Material online)

could well correspond to ancestral CNVs, because of adaptations of different populations in an ancestral species. It could even have participated in the speciation of these species similar to the situation in *Picea spp*. (Prunier et al. 2017).

Our results also showed the existence of correlation of *NUDX1-1a* activity not only with geraniol levels but also with some other volatiles (supplementary fig. S3, Supplementary Material online). This could be due *(i)* to an indirect effect (selection pressure on a transcription factor that regulates several biosynthesis genes, or pleiotropic effects), for example as it was observed for terpenes and phenylpropanoids in an overexpression experiment of *PAP1* in *R.* x *hybrida* 'Pariser Charme' (Ben Zvi et al. 2012), *(ii)* to diffuse selection pressure of pollinators, florivores, or parasites on several volatile compounds (for example, acyclic terpenoids and 2-phenylethanol are known to be very attractive for insects; Raguso 2004; Trhlin and Rajchard 2011), *(iii)* to common biosynthetic pathway for acyclic terpenoids, as it is the case in other species for geraniol, nerol, β-citronellol and their aldehydes and acetates for example (see review in Sun et al. 2016), or (*iv*) other unknown effects, like for example modifications or redirections of different fluxes through pathways of precursors or related to precursors.

In conclusion, *NUDX1* genes duplicated several times in *Rosaceae* species and probably acquired different functions. In the *Rosoidae* subfamily, three distinct clades were formed (fig. 10). The Nudx1-1 clade has evolved forming two subclades by duplication. In the genus *Rosa*, the more ancient *NUDX1-1b* gene was transposed from chromosome 4 and the surrounding TEs rearranged, such as the *Copia R24588* element, providing the building blocks for *box38*. This raises the question of how its promoter is specifically activated in the petals and by which transcription factors. The resulting *NUDX1-1a* on chromosome 2 was then able to produce geraniol in rose petals, which could be a high driving force of selection. This driving force was amplified in some rose species by several *cis*-duplications of *NUDX1-1a*. It is thus relevant to ask how the nonlinear effect of the gene copy number works in detail. Finally, use of the *box38* sequences for marker-assisted selection of scented roses could be a relevant application.

Materials and Methods

Plant materials and sampling

Samples (fig. 1c, and supplementary table S1, Supplementary Material online) were collected in France in several botanical gardens (Roseraie de Saint-Clair, Caluire, France; Roseraie de Loubert, Les Brettes, France; Parc de la Tête d'Or, Lyon, France), in the wild (Mornant, France), or in the BVpam laboratory garden (Saint-Etienne, France). The same species or variety in two different collections or different geographic area received two different names of accession. Descriptive data (ploidy, geography, phylogeny, and families) were reported according to the literature (Cairns 2003; Wissemann 2003; Schorr and Young 2007; Masure 2013; Fougère-Danezan et al. 2015; Zhu et al. 2015; Zhang et al. 2017; Debray et al. 2019). Each sampling was repeated at least three times between 2014 and 2019, depending on the location, the flowering period, and the weather forecast. This last point was important because wild roses often bloom during a fortnight. Buds for DNA extraction, and petals for mRNA

extraction, were frozen in liquid nitrogen for transport and conserved at -80°C before further experiments. Petals for volatile analysis were directly immersed in hexane containing (+/-)-camphor (#148075, Merck) at 5, 10, or 20 mg/l as an internal standard. Each vial contained 1 g of petals of individual flowers and 2 mL of hexane and (+/-)-camphor mix. Vials were transported to the laboratory in ice.

GC-MS analyses

The hexane extracts were recovered from the vial after 24 h at +4°C and processed according to (Sun et al. 2020): Agilent 6850 gas chromatograph, DB5 apolar capillary column (30 m x 0.25 mm), 7683B series injector, and 5973 Network mass selective detector (Agilent Technologies). Helium at a flow rate of 1.0 ml/min was used as a carrier gas with the following program: 40°C for 3 min, gradient of 3°C min from 40°C to 245°C, and 10 min at 245°C. Injection volume was 2 μl with a split mode (split ratio 1:2) and the injector and detector temperatures were 250°C. The parameters for mass spectrometer detector were set as follows: mass scan range 35 - 450 *m/z*, and ionization voltage 70 eV. Kovatz indexes (AI) were calculated according to Adams (2007) and to the Nist Web Book. Names and families of compounds (supplementary table S8, Supplementary Material online) were given by screening Wiley 275 and Nist 08 databases, and by names given by (Knudsen et al. 2006). Spearman's correlation coefficients (supplementary table S11, Supplementary Material online) and heatmap (supplementary fig. S3, Supplementary Material online) were calculated with the R language and environment (R Core Team 2015) using Hmisc (Harrell and Dupont 2020) and corrr (Kuhn et al. 2020) packages.

DNA and RNA extractions

For HMW-gDNA extraction, 100 mg of fresh buds were grinded with pestle and mortar in 2 ml of CTAB buffer (100 mM Tris-HCl pH 8.0, 3 M NaCl, 3% CTAB, 20 mM EDTA and 2% w/v PVP-40). 90 ng of Ribonuclease A (Sigma-Aldrich) was added before heating for 45 min in water bath at 65°C. Cellular debris were pelleted (13,000 x g, 5 min, 4°C) and the supernatant was mixed with equal volume of chloroform:isoamyl alcohol (24:1 v/v) and shaken slowly for 1 min. Aqueous phase was separated by centrifugation (12,000 x g, 5 min, 4°C). The upper phase was carefully recovered and washed 3 times more. Nucleic acids were precipitated by addition of 0.1 vol of 3 M sodium acetate pH 5.2 and 0.66 volume of cold ethanol 100% (-20˚C). Tubes were mixed by inversion and kept at -20˚C for 1 h. DNA was pelleted by centrifugation at 5,000 x g for 10 min at 4°C. DNA was washed 3 times with ethanol 70% and the pellet was dried for 10 min at room temperature and resuspended in 40 μl of TE (10 mM tris-HCL pH 8, 1 mM EDTA). All centrifugations were performed with slow acceleration and deceleration. Alternatively, the NucleoSpin® Plant II Kit was used (Macherey Nagel) for other experiment needing gDNA (cloning and qPCR).

For RNA extraction, petals of opened flowers (anthesis stage) were crushed in liquid nitrogen and extracted with the NucleoSpin® RNA Plant kit (Macherey-Nagel) with on-column DNAse for gDNA removal with the NucleoSpin® rDNAse Set (Macherey-Nagel). Absence of gDNA was checked by PCR. cDNA was obtained with the iScript Ready-to-use cDNA Supermix kit (Biorad) at 42°C for 1 h with 1 μg of RNA. All kits were used according to manufacturer's instructions.

qPCR, qRT-PCR and DNA cloning for sequencing

Primers used for cloning are given in supplementary table S7 (Supplementary Material online). Cloning of gDNAs and cDNAs (Clones_gDNAs_cDNAs.fasta, Supplementary Material online) were done after PCR amplification with Phusion High Fidelity polymerase (Thermo Fisher Scientific). The PCR parameters with RP7-FP7 primers were as followed: 98°C for 1 min, 28 cycles of [98°C for 10 sec, 58°C for 30 sec, and 72°C for 20 sec], and 72°C for 5 min. After purification of PCR product with the NucleoSpin® Gel and PCR clean up kit (Macherey-Nagel), ligation was done into pCRBlunt (Invitrogen), and transformed into *Escherichia coli* TOP10 (Invitrogen). Plasmid were purified with the NucleoSpin® Plasmid Kit (Macherey-Nagel). *NUDX1-1* gDNA and cDNA inserted into plasmids were sent to MWG Eurofins for sequencing using universal M13uni-21 primer.

Copy number determination of *NUDX1-1* genes by qPCR were performed with FP8-RP8 primers. The qPCR reaction consisted of 10 µl of SsoAdvancedTM SYBR Green Supermix (Bio-Rad), 500 nM R and F primers, 20 ng of diluted gDNA in 20 µl volume reaction. The parameters were as followed: 98°C for 5 min, 40 cycles of [98°C for 10 sec, and 58°C for 30 sec]. At the end of each run, the melting curve was set to 0.5°C every 2 sec from 65°C to 95°C. The number of copies was calculated by comparison with copies of *RcNUDX1-1* assuming that there were seven copies in Old Blush (five *RcNUDX1-1a* copies and two *RcNUDX1-1b* alleles; fig. 3, and supplementary fig. S5, Supplementary Material online). Three biological replicates were performed with gDNA from three different plants.

Amplifications for qRT-PCR were done according to (Sun et al. 2020) with housekeeping gene primers FP5-RP5 and FP6-RP6 designed on *RcEF1* and *RcTUB* sequences respectively (GenBank accession numbers BI978089, and AF394915) (Dubois et al. 2012). To determine the expression of the different *RcNUDX1-1* homologs of Old Blush, FP1-RP1 to FP4-RP4 primers were used. For *NUDX1-1* expression measurement in the different *Rosa* species, FP8-RP8 primers were used (fig. 7, supplementary fig. S5, and supplementary table S10, Supplementary Material online). Diluted (1/25) cDNAs were used in 20 µl reaction with SsoAdvancedTM SYBR Green Supermix (Bio-Rad). The PCR parameters were as followed: 95°C for 30 sec, and 30 cycles of [95°C for 5 sec, and 64°C for *RcEF1* amplification (GenBank accession number BI978089), or 58°C for *RcTUB* (GenBank accession number AF394915) and *NUDX1-1* amplification for 30 sec]. At the end of each run, the melting curve was set to 0.5°C every 2 sec from 65°C to 95°C. Cq values were automatically determined by the CFX96™ Real-Time system with default settings. ΔCt method (Pfaffl 2001) was used for quantification by comparison with reference genes. For each species, several independent qRT-PCR on different biological samples were performed.

Long read sequencing

Sequencing library was prepared from 1 µg fresh HMW-gDNA for each species using the genomic DNA ligation sequencing kit (SQK-LSK109, version 14aug2019, Oxford Nanopore Technologies) following manufacturer's recommandations. Library was then sequenced on a FLO-MIN106 flow cell using a MinION device (Oxford Nanopore Technologies). Obtained reads were subsequently basecalled using guppy software in high accuracy mode with parameters adapted to the sequencing kit and the flowcell [dna_r9.4.1_450bps_hac.cfg] using guppy in GPU mode. Basecalled fastq files were converted in fasta

using the fastq_to_fasta program from the FASTX Toolkit v0.0.14. Blast databases were obtained for each species from the fasta files then the blastn program (Camacho et al. 2009) was used to search for reads containing *NUDX* genes using either *RcNUDX1-1a, 1-1b, 1-2a, 1-2b, 1-2c,* and *1-3* sequences as query (supplementary tables S5, and S13, Supplementary Material online). Hits on identified reads were then manually analysed to determine the organisation of *NUDX* clusters.

Sequence annotations, phylogenies and synteny maps

Genes and transposons were named according to the GDR (Jung et al. 2019). The sequence of *R.* x *hybrida* cv. 'Papa Meilland' (*RhNUDX1*, GenBank accession number JQ820249) was used to clone the corresponding gene including the intron. It was named *RhNUDX1-rs* for reference sequence and was used to search sequences in "Rosa chinensis Genome v1.0 chromosomes" (Hibrand Saint-Oyant et al. 2018), "Rosa chinensis Old Blush Illumina Genome v1.0 chromosomes", "Rosa chinensis Old Blush homozygous Genome v2.0 chromosomes" (Raymond et al. 2018), "Rosa multiflora draft Genome v1.0" (Nakamura et al. 2017), "Fragaria vesca Genome v4.0" (Edger et al. 2018), "Malus x domestica Genome (GDDH13 v1-1)" (Daccord et al. 2017), and "Prunus persica Genome v2.0.a1" (Verde et al. 2013; Verde et al. 2017), all published in the GDR. They were searched directly using the blast tool online in the GDR, and/or by downloading the fasta files in Geneious Prime software (Biomatters Limited) for aligments, blastn, and calculation of identity. The non-assembled genome of *P. micrantha* "Potentilla micrantha v1.0" (Buti et al. 2018) of the GDR was also used because of the phylogeny proximity with the genus *Rosa*. Sequences were directly searched in its scaffolds by blastn in the Genious Prime software. The ML tree of fig. 2 was calculated and drawn in the Geneious Prime software with the plugin PhyML (Guindon et al. 2010) using complete DNA sequences, and non full-identical sequences. The following sequences published in Sun et al. (2020) were used as references to name clades: *RcNUDX1-1a* (*RcHm_v2.0_Chr2g0142071, 0142081, 0142111*,and *0142121*), *RcNUDX1-1b* (*RcHm_v2.0_Chr4g0436181*), *RcNUDX1-2a* (*RcHm_v2.0_Chr4g0436151*), *RcNUDX1-2b* (*RcHm_v2.0_Chr6g0244161*), *RcNUDX1-2c* (*RcHt_S2031.3*), *RcNUDX1-3* (*RcHm_v2.0_Chr4g0436191*), and *RwNUDX1-1, RwNUDX1-2a, RwNUDX1-2b, RwNUDX1-2c, RwNUDX1-2c', RwNUDX1-3* (Genbank accession numbers respectively MT362556 to MT362561). The gene sequences included the intron for increasing bootstraps (Align_Rosaceae_MLtree.fasta, and supplementary table S2, Supplementary Material online). *AtNUDX1* gene of *A. thaliana* was used as an outgroup (GenBank accession number AT1G68760). The dot-plot of similarity (supplementary fig. S4a, Supplementary Material online) was made with the plugin LASTZ (Harris 2007). For microsynteny (figs. 3, 4, 6, and supplementary table S2, Supplementary Material online), marker genes around the *NUDX1* genes were used to verify correspondences between homologous regions in the GDR and in MinION reads. They were arbitrarily named A to S (full list in supplementary table S14, Supplementary Material online).

The *NUDX1* gene phylogeny (fig. 5, supplementary fig. S2, and Clones_IntronExonStructure.fasta, Supplementary Material online) was reconstructed using the entire 660 bp, thus including the intron, with *F. vesca NUDX1* gene as outgroup (GenBank accession number XM_004297107.2). Indeed, as the

coding parts of the *NUDX1* gene are strongly conserved between species, too little phylogenetic information is contained in the exonic sequences, while the intronic sequence is more variable and makes the phylogenetic reconstruction possible. *NUDX1* genes were aligned using Clustalw (Thompson et al. 2002), and sites ambiguously aligned were removed with Gblocks (Castresana 2000), resulting in a 608 bp alignment. ML phylogenetic reconstruction was conducted using PhyML (Guindon et al. 2010) under a GTR+G+I model (Align_OldBlush_MLtree.fasta, Supplementary Material online). Tree was rooted with the *FvNUDX1-1* gene (GenBank accession number XM_004297107.2). In order to understand the history of duplication, we need to know which sequences belongs to the chromosome 2 (*NUDX1-1a*) paralog and which ones belong to the chromosome 4 (*NUDX1-1b*) paralog. To achieve that, all sequences were aligned by blastn against Old Blush *RcNUDX1-1a* (GenBank accession number, CM009583.1, from position 59,567,055 to 59,567,676 bp) and *RcNUDX1-1b* (GenBank accession number CM009585.1, from position 59,520,245 to 59,520,862 bp). Identities of the DNA sequences and the putative proteins were also calculated (supplementary tables S3 and S4, Align_OldBlush_DNAsequences.fasta, and Align_OldBlush_Proteins.fasta, Supplementary Material online) to draw the comprehensive map (fig. 3). gDNAs displaying identity more than 1% higher with *RcNUDX1-1a* than with *RcNUDX1-1b* were assigned to the Nudx1-1a subclade and *vice versa* (supplementary fig. S2, and supplementary table S9, Supplementary Material online). As these two paralogs are very similar, some sequences aligned similarly with blastn (less than 1% with both references), and thus were not assigned to one of the subclades.

Promoter analysis, cloning and transient expression

For promoter analyses of *Copia R24588* and *box38* hits and homology, we used blastn (Camacho et al. 2009) with the minimum seed size [word_size = 7] allowing to recover hits from short query sequences. Multiple aligments were performed with MAFFT (Kato et al. 2019) using the following parameters [parameters --thread 2 --reorder --adjustdirectionaccurately --anysymbol --maxiterate 2 --retree 1 –genafpair]. Aligments are given in Align_CopiaBox38_Chr2.fasta and Align_CopiaLTR_Chr2and4.fasta (Supplemental Materials online). Quality control of the alignment and minor extensions of the blastn hits (up to two bp) within the *box38* consensus were performed manually. A consensus sequence logo for *box38* was created using WebLogo v2.8.2 (Crooks et al., 2004). We also mapped the consensus sequence of *Copia R24588* of the GDR by using RepeatClassifier, a tool included with RepeatModeler2 (Flynn et al. 2020) and TE-Aid (https://github.com/clemgoub/TE-Aid).

Primers used for cloning are given in supplementary table S7 (Supplementary Material online). For promoter cloning, FP9-RP9 (upstream region of *NUDX1-1b*) and FP10-RP10 (upstream region of *NUDX1-1a*) were used and cloned into pCRBlunt (Invitrogen) as mentioned above and sequenced with the same procedure using the M13uni-21 primer for sequencing. Amplification of the different promoter regions was done with Phusion U Hot Start DNA Polymerase (Thermo Fisher Scientific) with combinations of USER extended primer FP11 to FP15 and RP11 with RcOB gDNA as template (fig. 9). The PCR parameters primers were as followed: 98°C for 1 min, 25 cycles of [98°C for 10 sec, 60°C for 30 sec, and 72°C for 30 sec], and 72°C for 5 min. PCR products were cloned into a pCAMBIA2300

binary base vector with linearized *Pac*I-USER cassette upstream the GFP and NOS-terminator using USER enzyme (New England Biolabs). The control construct based on double CaMV *35S* promoter was cloned into the same vector with the same method using the binary vector pMDC32 containing this promoter as matrix with FP16-RP16. All USER reaction was transformed into *E. coli* TOP10 (Invitrogen). Plasmids were purified with the NucleoSpin plasmid kit (Macherey Nagel). Sequence of constructs were verified before use.

These constructs were transformed into the *Agrobacterium* strain LBA4404. Agrobacteria were grown on LB agar with rifampicin (50 μg/ml), gentamicin (20 μg/ml), and kanamycin (50 μg/ml), and then screened by PCR for the presence of the construct. Agrobacteria were grown in 25 ml of liquid LB with antibiotics and collected by centrifugation at room temperature for 8 min at 4,000 x g and washed in 10 mM $MgCl_2$ and 10 mM MES pH 5.7 buffer 3 times. They were diluted to OD_{600nm} = 1.0 with wash buffer and infiltrated on the abaxial side of Old Blush petals with a syringe. After 3 days, infiltrated petals were observed with a TCS-SP2 inverted confocal scanning laser microscope (Leica) with a x40/0.80W lens. The argon laser was set at 488 nm for GFP excitation and the fluorescent signal was captured at 500 to 550 nm.

Enzyme assay

RcNUDX1-1a and *RcNUDX1-1b* cDNA sequences corresponding to Old Blush gDNA1 and 2, respectively (Clones_gDNAS_cDNAs.fasta, Supplementary Material online), were amplified by PCR (primers FP17-RP17; supplementary table S7, Supplementary Material online) and cloned in pET-30a(+) between the *Kpn*I and *Sal*I restriction sites. *RmNUDX1-1a* and *RmNUDX1-1b* cDNAs corresponding to Moschata gDNA10 and gDNA2, respectively (Clones_gDNAS_cDNAs.fasta, Supplementary Material online), were synthetized (GenScript) and cloned in pET-30a(+) between the *Kpn*I and *Sal*I restriction sites. Sequences and vectors were verified by sequencing and transformed into *E. coli* BL21(DE3)pLysS.

Transformants were grown at 37°C in LB medium until OD_{600nm} = 0.4. Proteins were produced by overnight induction at 16°C with 1 mM IPTG. After centrifugation, bacteria pellet was resuspended in buffer (50 mM Tris-HCL pH 8.5, 500 mM NaCl, 2 mM DTT, 8% glycerol v/v, 10 mM imidazole, 0.25 mg ml lysozyme) and lysed by sonication. Supernatant was mixed with Ni-NTA agarose resin (Qiagen) for 1 h. Resin was rinsed 5 times with 50 mM Tris-HCL pH 8.5, 500 mM NaCl, 2 mM DTT, 8% v/v glycerol, and 50 mM imidazole, and finally eluted in the same buffer but containing 250 mM imidazole. Proteins were desalted by passing through a PD10 desalting column (GE Healthcare) equilibrated with the assay buffer (50 mM HEPES pH 8, 5 mM $MgCl_2$, 5% v/v glycerol) and quantified with the Bradford method. All steps of purification were conducted on ice.

Enzymatic reactions were performed in assay buffer containing different concentrations of GPP (0.5, 1, 2, 5, 10 ,30 or 50 μM) in 100 μl reaction volume at 30°C for 4 min, and using 20 ng of proteins. Reactions were stopped by adding 100 μL MeOH:H_2O (10 mM NH_4OH) 7:3 and mixed for 30 s. Product analysis were performed on an Agilent 1260 infinity II LC system coupled to an Agilent Ultivo triple quadrupole mass spectrometer (Agilent Technologies, Santa Clara, USA) using a Poroshell 120

HPH-C18 column (50 mm x 2.1 mm, particle size 1.9 μm, Agilent) heated at 35°C. The mobile phases consisted of 10 mM ammonium bicarbonate pH 10.2 with 0.15% v/v ammonia, as solvent A, and acetonitrile with 0.15 % v/v ammonia, as solvent B, with a 0.6 mL min flow rate. 2 μl of reaction mixture was injected for each sample. Separation was achieved with a gradient starting with 2% B reaching 98 % B in 2 min, 1 min isocratic at 98 % B and return at 2 % B at 3.10 min with equilibration until 6.5 min. Mass spectrometer tunings were as follow: capillary voltage 5000 V, gas temperature 350 °C, gas flow 12 L/min, and nebulizer 55 psi. Products detection was achieved in negative and MRM modes with the following MS/MS transitions and tunings: 312.2 to 78.9 *m/z* for GPP with Fragmentor at 70 V and Collision Energy at 92 V and 233.1 to 78.9 *m/z* for GP with Fragmentor at 75 V, and collision energy at 60 V. Data analysis was performed with MassHunter quantitative software (Agilent Technologies). Enzyme Kinetic parameters were determined using the Lineweaver-Burk plot model.

Acknowledgments

This work was supported by Agence Nationale de la Recherche (grant number ANR-16-CE20-0024-01 to S.B.), by Centre National de la Recherche Scientifique (grant reference MITI-ExoMod to J.C.C.), by Fondation de l'Université Jean Monnet (grant to J.C.C.), by National Science Foundation IOS (grant number 1655438 to N.D.), and by USDA National Institute of Food and Agriculture Hatch Project (grant number 177845 to N.D.). The authors wish to thank Thérèse Loubert and the town halls of Caluire and of Lyon, for sampling authorization in "Roseraie de Loubert", "Roseraie de Saint-Clair", and "Parc de la Tête d'Or". They also thank "Groupe de Recherche MédiatEC", and "Réseau MétaSP" for the discussions on specialized metabolism, and Laurent Duret and Tristan Lefébure (laboratoire de Biométrie et Biologie Evolutive, Lyon, France) for the discussions on gene evolution and functionalization.

Data availability

Raw data are given in Supplementary Material online, including fasta sequences of cDNAs and gDNAs cloned in this paper. *NUDX1-rs* sequence is deposited in the GenBank with the accession number MW762674. Reads from MinION sequencing are available in the SRA database in FASTQ format under the bioproject accession number PRJNA706580.

References

Axelsson E, Ratnakumar A, Arendt ML, Maqbool K, Webster MT, Perloski M, Liberg O, Arnemo JM, Hedhammar A, Lindblad-Toh K. 2013. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature*. 495:360-364.

Baudino S, Hugueney P, Caissard JC. 2020. Evolution of scent genes. In: Pichersky E, Dudareva N, editors. Biology of plant volatiles. Boca Raton: CRC Presse. p. 217-234.

Ben Zvi MM, Shklarman E, Masci T, Kalev H, Debener T, Shafir S, Ovadis M, Vainstein A. 2012. PAP1 transcription factor enhances production of phenylpropanoid and terpenoid scent compounds in rose flowers. *New Phytol*. 195:335-345.

Boutanaev AM, Osbourn AE. 2018. Multigenome analysis implicates miniature inverted-repeat transposable elements (MITEs) in metabolic diversification in eudicots. *Proc. Natl Acad. Sci. USA.* 115: E6650-E6658.

Buti M, Moretto M, Barghini E, Mascagni F, Natali L, Brilli M, Lomsadze A, Sonego P, Giongo L, Alonge M, et al. 2018. The genome sequence and transcriptome of Potentilla micrantha and their comparison to *Fragaria vesca* (the woodland strawberry). *Gigascience*. 7:1-14.

Cairns T. 2003. Classification. In: Roberts A, Debener T, Gudin S, editors. Encyclopedia of Rose Science Amsterdam, The Netherlands: Elsevier. p. 117-123.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics*. 10: 421.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*. 17:540-552.

Cerbin S, Jiang N. 2018. Duplication of host genes by transposable elements. *Curr Opin Genet Dev*. 49:63-69.

Chen F, Tholl D, D'Auria JC, Farooq A, Pichersky E, Gershenzon J. 2003. Biosynthesis and emission of terpenoid volatiles from Arabidopsis flowers. *Plant Cell*. 15:481-494.

Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo : a sequence logo generator. *Genome Res.* 14, 1188-1190.

Daccord N, Celton J-M, Linsmith G, Becker C, Choisne N, Schijlen E, van de Geest H, Bianco L, Micheletti D, Velasco R, et al. 2017. High-quality *de novo* assembly of the apple genome and methylome dynamics of early fruit development. *Nat Genet*. 49:1099-1106.

DeBolt S. 2010. Copy Number Variation shapes genome diversity in Arabidopsis over immediate family generational scales. *Genome Biol Evol*. 2:441-453.

Debray K, Marie-Magdelaine J, Ruttink T, Clotault J, Foucher F, Malécot V. 2019. Identification and assessment of variable single-copy orthologous (SCO) nuclear loci for low-level phylogenomics: a case study in the genus *Rosa* (Rosaceae). *BMC Evol Biol*. 19:152.

Dubois A, Carrere S, Raymond O, Pouvreau B, Cottret L, Roccia A, Onesto JP, Sakr S, Atanassova R, Baudino S, et al. 2012. Transcriptome database resource and gene expression atlas for the rose. *BMC Genomics*. 13:638-648.

Edger PP, VanBuren R, Colle M, Poorten TJ, Wai CM, Niederhuth CE, Alger EI, Ou S, Acharya CB, Wang J, et al. 2018. Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (*Fragaria vesca*) with chromosome-scale contiguity. *Gigascience*. 7:1-7.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF, 2020. RepeatModeler2 for automated genomic discovery of transposable elements families. *Proc Natl Acad Sci USA.* 117, 9451-9457.

Fougère-Danezan M, Joly S, Bruneau A, Gao XF, Zhang LB. 2015. Phylogeny and biogeography of wild roses with specific attention to polyploids. *Ann Bot*. 115:275-291.

Galindo-González L, Mhiri C, Deyholos MK, Grandbastien MA. 2017. LTR-retrotransposons in plants: Engines of evolution. *Gene*. 626:14-25.

Grandbastien M-A. 2015. LTR retrotransposons, handy hitchhikers of plant regulation and stress response. *Biochem Biophys Acta*. 1849: 403-416.

Gu T, Han Y, Huang R, McAvoy RJ, Li Y. 2016. Identification and characterization of histone lysine methylation modifiers in *Fragaria vesca*. *Sci Rep*. 6:23581.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 59:307-321.

Hahn MW. 2009. Distinguishing among evolutionary models for the maintenance of gene duplicates. *J Hered*. 100:605-617.

Han Y, Gasic K, Korban SS. 2007. Multiple-Copy Cluster-Type organization and evolution of genes Encoding *O*-Methyltransferases in the Apple. *Genetics*. 176:2625-2635.

Han Y, Korban SS. 2007. *Spring*: A novel family of miniature inverted-repeat transposable elements is associated with genes in apple. *Genomics*. 90:195-200.

Harrell F, Dupont C. 2020. Hmisc: Harrell miscellanous R package version 4.4.2.

Harris RS. 2007. Improved paiwise alignment of genomic DNA. [Ph.D. thesis]: Pennsylvania State University.

Henry LK, Gutensohn M, Thomas ST, Noel JP, Dudareva N. 2015. Orthologs of the archaeal isopentenyl phosphate kinase regulate terpenoid production in plants. *Proc Natl Acad Sci USA*. 112:10050-10055.

Henry LK, Thomas ST, Widhalm JR, Lynch JH, Davis TC, Kessler SA, Bohlmann J, Noel JP, Dudareva N. 2018. Contribution of isopentenyl phosphate to plant terpenoid metabolism. *Nat Plants*. 4:721-729.

Hibrand Saint-Oyant L, Ruttink T, Hamama L, Kirov I, Lakhwani D, Zhou NN, Bourke PM, Daccord N, Leus L, Schulz D, et al. 2018. A high-quality genome sequence of *Rosa chinensis* to elucidate ornamental traits. *Nat Plants*. 4:473-484.

Iwata H, Gaston A, Remay A, Thouroude T, Jeauffre J, Kawamura K, Oyant LH-S, Araki T, Denoyes B, Foucher F. 2012. The *TFL1* homologue *KSN* is a regulator of continuous flowering in rose and strawberry. *Plant J*. 69:116-125.

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature*. 431:569-573.

Jiang S, Wang X, Shi C, Luo J. 2019. Genome-Wide identification and analysis of High-Copy-Number LTR retrotransposons in asian pears. *Genes*. 10:156.

Jung S, Lee T, Cheng C-H, Buble K, Zheng P, Yu J, Humann J, Ficklin SP, Gasic K, Scott K, et al. 2019. 15 years of GDR: New data and functionality in the Genome Database for Rosaceae. *Nucleic Acids Res*. 47:D1137-D1145.

Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: multiple sequence aligment, interactive sequence choice ad visualization. *Brief Bioinformatics.* 20, 1160-1166.

Knudsen JT, Eriksson R, Gershenzon J, Ståhl B. 2006. Diversity and distribution of floral scent. *Bot Rev*. 72:1-120.

Kobayashi S, Goto-Yamamoto N, Hirochika H. 2004. Retrotransposon-induced mutations in grape skin color. *Science.* 304: 982.

Krasileva KV. 2019. The role of transposable elements and DNA damage repair mechanisms in gene duplications and gene fusions in plant genomes. *Curr Opin Plant Biol*. 48:18-25.

Kuhn M, Jackson s, Cimentada J. 2020. Corrr: correlations in R. R package version 0.4.3.

Lescot M, Déhais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, Rouzé P, Rombauts S. 2002. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res*. 30:325-327.

Li W, Lybrand DB, Xu H, Zhou F, Last RL, Pichersky E. 2020. A Trichome-Specific, Plastid-Localized *Tanacetum cinerariifolium* Nudix Protein Hydrolyzes the Natural Pyrethrin Pesticide Biosynthetic Intermediate *trans*-Chrysanthemyl Diphosphate. *Front Plant Sci*. 11:482.

Lu H, Giordano F, Ning Z. 2016. Oxford Nanopore MinION sequencing and genome assembly. *Genomics Proteomics Bioinformatics.* 14: 265-279.

Lye ZN, Purugganan MD. 2019. Copy Number Variation in Domestication. *Trends Plant Sci*. 24:352-365.

Magnard J-L, Roccia A, Caissard J-C, Vergne P, Sun P, Hecquet R, Dubois A, Hibrand-Saint Oyant L, Jullien F, Nicolè F, et al. 2015. Biosynthesis of monoterpene scent compounds in roses. *Science*. 349:81-83.

Maron LG, Guimarães CT, Kirst M, Albert PS, Birchler JA, Bradbury PJ, Buckler ES, Coluccio AE, Danilova TV, Kudrna D, et al. 2013. Aluminum tolerance in maize is associated with higher *MATE1* gene copy number. *Proc Natl Acad Sci USA*. 110:5241-5246.

Masure P. 2013. Guide des rosiers sauvages. Paris, France: Delachaux et Niestlé.

McLennan A. 2013. Substrate ambiguity among the nudix hydrolases: biologically significant, evolutionary remnant, or both? *Cell Mol Life Sci*. 70:373-385.

Morata J, Marín F, Payet J, Casacuberta JM. 2018. Plant lineage-specific amplification of Transcription Factor Binding Motifs by Miniature Inverted-Repeat Transposable Elements (MITEs). *Genome Biol Evol*. 10:1210-1220.

Nakamura N, Hirakawa H, Sato S, Otagaki S, Matsumoto S, Tabata S, Tanaka Y. 2017. Genome structure of *Rosa multiflora*, a wild ancestor of cultivated roses. *DNA Res*. 25:113-121.

Ono K, Akagi T, Morimoto T, Wünsch A, Tao R. 2018. Genome re-sequencing of diverse sweet cherry (*Prunus avium*) individuals reveals a modifier gene mutation conferring pollen-part self-compatibility. *Plant Cell Physiol*. 59:1265-1275.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat Genet*. 39:1256-1260.

Pfaffl MW. 2001. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res*. 29:e45.

Prunier J, Caron S, MacKay J. 2017. CNVs into the wild: screening the genomes of conifer trees (*Picea spp.*) reveals fewer gene copy number variations in hybrids and links to adaptation. *BMC Genomics*. 18:97.

R Core Team. 2015. R: A language and environment for statistical computing Vienna, Austria.

Raguso RA. 2004. Why do flowers smell? The chemical ecology of fragrance-driven pollination. In: Millar JG, Cardé RT, editors. Advances in Insect Chemical Ecology. Cambridge: Cambridge University Press. p. 151-178.

Raymond O, Gouzy J, Just J, Badouin H, Verdenaud M, Lemainque A, Vergne P, Moja S, Choisne N, Pont C, et al. 2018. The *Rosa* genome provides new insights into the domestication of modern roses. *Nat Genet*. 50:772-777.

Roccia A, Hibrand-Saint Oyant L, Cavel E, Caissard J-C, Machenaud J, Thouroude T, Jeauffre J, Bony A, Dubois A, Vergne P, et al. 2019. Biosynthesis of 2-phenylethanol in rose petals is linked to the expression of one allele of *RhPAAS*. *Plant Physiol.* 179: 1064-1079.

Schorr P, Young MA. 2007. Modern Roses 12: The Comprehensive List of Roses in Cultivation Or of Historical Or Botanical Importance: American Rose Society.

Shirai K, Hanada K. 2019. Contribution of functional divergence through copy number variations to the inter-species and intra-species diversity in specialized metabolites. *Front Plant Sci*. 10:1567.

Srouji JR, Xu A, Park A, Kirsch JF, Brenner SE. 2017. The evolution of function within the Nudix homology clan. *Proteins*. 85:775-811.

Sun P, Dégut C, Réty S, Caissard J-C, Hibrand-Saint Oyant L, Bony A, Paramita SN, Conart C, Magnard J-L, Jeauffre J, et al. 2020. Functional diversification in the Nudix hydrolase gene family drives sesquiterpene biosynthesis in *Rosa* x *wichurana*. *Plant J*. 104:185-199.

Sun P, Schuurink RC, Caissard JC, Hugueney P, Baudino S. 2016. My Way: Noncanonical biosynthesis pathways for plant volatiles. *Trends Plant Sci*. 21:884-894.

Thompson JD, Gibson TJ, Higgins DG. 2002. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics*. Chapter 2:Unit 2.3.

Trhlin M, Rajchard J. 2011. Chemical communication in the honeybee (*Apis mellifera* L.): a review. *Veterinarni Medicina*. 56:265-273.

Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, Marroni F, Zhebentyayeva T, Dettori MT, Grimwood J, Cattonaro F, et al. 2013. The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat Genet*. 45:487-494.

Verde I, Jenkins J, Dondini L, Micali S, Pagliarani G, Vendramin E, Paris R, Aramini V, Gazza L, Rossini L, et al. 2017. The Peach v2.0 release: high-resolution linkage mapping and deep resequencing improve chromosome-scale assembly and contiguity. *BMC Genomics*. 18:225.

Wang A, Yamakake J, Kudo H, Wakasa Y, Hatsuyama Y, Igarashi M, Kasai A, Li T, Harada T. 2009. Null mutation of the *MdACS3* gene, coding for a ripening-specific 1-aminocyclopropane-1-carboxylate synthase, leads to long shelf life in apple fruit. *Plant Physiol*. 151:391-399.

Wang L, Peng Q, Zhao J, Ren F, Zhou H, Wang W, Liao L, Owiti A, Jiang Q, Han Y. 2016. Evolutionary origin of Rosaceae-specific active non-autonomous *hAT* elements and their contribution to gene regulation and genomic structural variation. *Plant Mol Biol*. 91:179-191.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*. 8:973-982.

Wissemann V. 2003. Conventional taxonomy of wild roses. In: Roberts A, Debener T, Gudin S, editors. Encyclopedia of rose science. London: Elsevier. p. 111-117.

Xiang Y, Huang CH, Hu Y, Wen J, Li S, Yi T, Chen H, Xiang J, Ma H. 2017. Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Mol Biol Evol*. 34:262-281.

Yoshimura K, Shigeoka S. 2015. Versatile physiological functions of the Nudix hydrolase family in Arabidopsis. *Biosci Biotechnol Biochem*. 79:354-366.

Zhang L, Hu J, Han X, Li J, Gao Y, Richards CM, Zhang C, Tian Y, Liu G, Gul H, et al. 2019. A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. *Nat Commun*. 10:1494.

Zhang S-D, Jin J-J, Chen S-Y, Chase MW, Soltis DE, Li H-T, Yang J-B, Li D-Z, Yi T-S. 2017. Diversification of Rosaceae since the Late Cretaceous based on plastid phylogenomics. *New Phytol*. 214:1355-1367.

Zhao D, Ferguson AA, Jiang N. 2016. What makes up plant genomes: The vanishing line between transposable elements and genes. *Biochim Biophys Acta*. 1859:366-380.

Zhu ZM, Gao XF, Fougere-Danezan M. 2015. Phylogeny of Rosa sections Chinenses and Synstylae (Rosaceae) based on chloroplast and nuclear markers. *Mol Phylogenet Evol*. 87:50-64.

Żmieńko A, Samelak A, Kozłowski P, Figlerowicz M. 2014. Copy number polymorphism in plant genomes. *Theor Appl Genet*. 127:1-18.

Table 1. Comparison of geraniol concentration and expression of *NUDX1-1* homologs in wild and heritage roses.

Accession names[a]	Geraniol concentration (*μ*g/gFW)	qRT-PCR on *NUDX1-1* homologs (a.u.)[b]	Number of cDNA clones[c]	Number of gDNA[d] clones[c]
Arvensis_B	0.0 (0.0)[e]	0.1 (0.1)[e]	0	2

Banksiae	0.0 (0.0)	0.0 (0.0)	0	4
Bracteata	0.0 (0.0)	0.0 (0.0)	0	1
Chinensis	0.0 (0.0)	0.0 (0.0)	0	2
Gigantea	0.0 (0.0)	0.0 (0.0)	0	2
Laevigata	0.0 (0.0)	0.0 (0.0)	0	1
Mirifica	0.0 (0.0)	0.0 (0.0)	0	5
Roxburghii	0.0 (0.0)	0.0 (0.0)	0	4
Rubus	0.0 (0.0)	0.7 (0.0)	3	6
Sericea	0.8 (0.3)	0.0 (0.0)	0	6
Foetida	3.0 (2.1)	13.3 (12.6)	2	3
Persian_Yellow	5.4 (1.9)	34.8 (30.3)	1	6
Ecae	5.8 (0.2)	0.0 (0.0)	1	3
Hugonis_B	17.9 (2.3)	0.0 (0.0)	nd[f]	2
Canina	22.9 (6.0)	111.5 (1.6)	1	15
Phoenicia	27.0 (4.4)	155.2 (12.2)	1	7
Moschata	29.5 (10.2)	111.6 (5.8)	3	11
Fedtschenkoana	37.8 (5.9)	87.2 (3.3)	1	6
Rugosa	44.4 (24.2)	36.1 (24.0)	1	12
Centifolia	45.3 (17.2)	207.1 (131.9)	3	14
Arvensis_A	53.9 (52.8)	256.8 (139.7)	2	5
Gallica_B	63.4 (3.9)	91.5 (20.2)	1	3
Autumn_Damask	84.1 (11.6)	63.1 (18.8)	3	7
Hugonis_A	89.8 (31.4)	12.5 (0.3)	nd	4
Nutkana	96.3 (26.3)	374.1 (129.2)	2	6
Old_Blush	99.8 (5.9)	61.0 (12.0)	1	2
Pendulina	104.4 (45.4)	174.8 (82.9)	2	3
Villosa	107.9 (10.4)	128.0 (6.0)	1	5
Gallica_A	108.7 (11.6)	88.3 (21.7)	2	6
Damask_Kazanlik	112.2 (39.5)	43.1 (1.6)	3	9
Majalis	112.6 (2.0)	25.3 (1.8)	2	7
Carolina	145.7 (40.7)	339.4 (100.7)	3	6
Woodsii	180.4 (2.8)	19.8 (3.0)	2	5
Officinalis	192.1 (42.5)	112.5 (8.3)	1	5
Spinosissima	nd	nd	1	14

[a] For the rose accession names see supplementary table S1, Supplementary Material online.

[b] Amplification with FP8-RP8 primers (supplementary table S7, Supplementary Material online).

[c] Cloning with FP7-RP7 primers (supplementary table S7, Supplementary Material online).

[d] The number of gDNA clones correspond to different genomic sequences from ATG to STOP codons (supplementary table S9, and Clones_gDNAs_cDNAs.fasta, Supplementary Material online). They all included a single intron (Clones_IntronExonStructure.fasta, Supplementary Material online).

[e] Values correspond to averages, and SD are given in parentheses. Extensive values are given in supplemental tables S8 and S10 (Supplementary Material online).

[f] Not done.

Figure legends

Fig. 1. Overview of the evolution of the *Rosaceae* family and of the *Rosa* genus.

a. Antique murals in Knossos (approx. 1 700 B.C.). Arrow show the original drawing of a wild rose (the other drawing was made during an irreversible restoration).

b. Antique murals in Pompei (approx. 79 A.C.). Roses were painted with dozens of petals (arrow).

c. Synthetic phylogeny and evolution diagram obtained by simplification of data from (Fougère-Danezan et al. 2015; Zhu et al. 2015; Xiang et al. 2017; Zhang et al. 2017, and (Debray et al. 2019). Only species and varieties used or cited in our article are shown (supplementary table S1, Supplementary Material online). Some species are written in grey because their phylogenetic position is discussed (*R. moschata, R. rugosa*), or because they are allopolyploids (*R. canina, R. spinosissima*). *R. foetida* and *R. stellata mirifica* are not shown because of their unresolved position. Heritage roses also include some crosses made by breeders, which are not considered as botanical roses, and which are not shown here.

Fig. 2. ML tree of genomic sequences of *NUDX1* homologs in the Rosaceae.

The tree was made with sequences of (Sun et al. 2020), and with sequences obtained by blastn (from ATG to STOP including the intron) in selected species of the GDR (Align_Rosaceae_MLtree.fasta, Supplementary Material online). *AtNUDX1* gene was used to root the tree (large black arrow). *RhNUDX1-rs* was added for reference (large orange arrow). Clades were named according to (Sun et al. 2020). Numbers correspond to bootstraps (%). Scale bar represent substitution per site.

Fig. 3. Gene map of *RcNUDX1* in Old Blush.

Each pair of homologous chromosomes are shown. Similar regions including *RcNUDX1* sequences are highlighted in grey between the two homologous chromosomes. Gene lengths, from the ATG codon to the STOP, including introns, and intergenic lengths are indicated. However, the picture does not respect these lengths. Gene numbers were obtained by making a systematic inventory of chromosomes on the three genomes of Old Blush published in the GDR and by comparison with our MinION long reads (supplementary tables S2, S5, and Align_OldBlush_DNAsequences.fasta, Supplementary Material online), but only sequence accessions useful for mapping are shown. Null alleles were confirmed on chromosomes 2 and 7 because scaffolds available in the GDR including both upstream and downstream regions were found. All null alleles were also confirmed by MinION sequencing (supplementary table S5, Supplementary Material online).

Large orange arrows, genes from Nudx1-1 clade; large blue arrows, genes from Nudx1-2 clade; large green arrows, genes from Nudx1-3 clade. Copies of *RcNUDX1-1a* are arbitrarily numbered in orange on chromosome 2. Sequences with a dashed outline are pseudogenes including STOP codons. Chr, chromosomes. Marker genes (grey arrows) used for microsynteny are listed in supplementary table S14 (Supplementary Material online). On chromosome 2, gene D was not found on scaffold *RcHt_S929* but useful in MinION reads. On chromosome 6, marker genes were not found around the null allele in the GDR, but MinION long reads included marker genes J, K, L, and *RcNUDX1-2b,* or its null allele (read numbers in supplementary Table S5, Supplementary Material online).

Fig. 4. Synteny map of the *Rosaceae* genomes.

a. Microsynteny of chromosome 4 of Old Blush in the cluster region of *RcNUDX1-1b, RcNUDX1-2a* and *RcNUDX1-3*.

b. Microsynteny of chromosome 2 of Old Blush in the cluster region of *RcNUDX1-1a.*

Chromosome numbers are indicated except for *P. micrantha* for which the genome was non-assembled in the GDR (supplementary table S2, Supplementary Material online). Large orange arrows, genes from Nudx1-1 clade; large blue arrows, genes from Nudx1-2 clade; large green arrows, genes from Nudx1-3 clade; large violet arrows, sequences of the Nudx1-4 clade; large black arrows, other *NUDX1* genes; large white arrows, unique genes; large grey arrows, genes used for microsynteny (marker genes are listed in supplementary table S14, Supplementary Material online). Accession numbers of *NUDX1* genes are in supplementary table S2 (Supplementary Material online). There was no sequence of *NUDX1* in the microsyntenic regions of chromosomes 6 and 7. Distances between sequences and scales are approximative, and gene lengths and TE sizes are distorted to show the relative organization. Chr, chromosomes.

Fig. 5. ML tree of genomic sequences of the Nudx1-1 clade.

Orange asterisks indicate species in which a cDNA clone is the exact ORF of the gDNA (supplementary table S9, Clones_IntronExonStructure.fasta, Supplementary Material online). Blue stars indicate species not producing geraniol (table 1, and supplementary table S8, Supplementary Material online). Large orange and red arrows indicate respectively the *RcNUDX1-1a* and *RcNUDX1-1b* genes of Old Blush. White dots correspond to bootstraps less than 70%, grey dots, between 70 and 95%, and black dots, more than 95%. The tree is rooted with a sequence of *F. vesca* (large black arrow). For the extended tree see supplementary fig. S2 and Align_OldBlush_MLtree.fasta (Supplementary Material online).

Fig. 6. Organization of the shared TEs around the *NUDX1-1a* and *NUDX1-1b* sequences in three accessions: Old Blush, Moschata, and Laevigata.

a. Chromosome 2 of Old Blush and corresponding microsyntenic regions of Moschata and Laevigata accessions. The cluster could be interpreted with two types of putative blocks (show on a top), which could then duplicate into five blocks. In the first hypothesis, MITEs are missing in block #5. In the second hypothesis, MITEs are missing in block #1.

b. Chromosome 4 of Old Blush and corresponding microsyntenic regions of Moschata and Laevigata accessions (MinION sequencing in supplementary table S13, Supplementary Material online). Only shared TEs are shown (supplementary table S12, Supplementary Material online).
Large orange arrows, genes from Nudx1-1 clade; large blue arrows, genes from Nudx1-2 clade; large green arrows, genes of Nudx1-3 clade; pink triangles, MITE *P580.2030*; dark blue triangles, MITE *G13554*; yellow arrow, *Copia R24588*; large grey arrows, marker genes used to find reads in the MinION database (supplementary table S14, Supplementary Material online). Distances between sequences are approximate and gene lengths and TE sizes are distorted to show the relative organization. Chr, chromosomes.

Fig. 7. Correlation between the expression of *NUDX1-1* homologs and the number of gene sequences in rose species.
Expression of *NUDX1-1* was determined by qRT-PCR with FP8-RP8 primers, and FP5-RP5 and FP6-RP6 primers for reference genes (supplementary tables S7, and S10, Supplementary Material online). Number of gene sequences was estimated by qPCR with FP8-RP8 primers (supplementary fig. S5, Supplementary Material online). Error bars correspond to SD. a.u., arbitrary units.

Fig. 8. Alignment interpretation of *box38* of chromosomes 2 and 4 of Old Blush genome.
a. An interpretative map of a block on chromosome 2 showing the localization of *Copia R24588* and *box38* A fragment in the promoter of *RcNUDX1-1a*.
b. Manual annotation of *Copia R24588* consensus with the different regions of the retrotransposon.
c. Alignment (MAAFT) of *Copia R24588* consensus and upstream regions of *RcNUDX1-1a* on chromosome 2, and $^{\Psi}$*RcNUDX1-2a* on chromosome 4 (alignment is given in Align_CopiaLTR_Chr2and4.fasta, Supplemental Material online). This *Copia R24588* fragment aligns 4 bp further with the *box38* consensus (37/38 bp) than the fragments seen in the repeat blocks of chromosome 2, strengthening the LTR origin hypothesis for *box38*.
Red circle, *Copia R24588* consensus of the GDR (Jung et al. 2019; Raymond et al. 2018); Brown circle, upstream region of $^{\Psi}$*RcNUDX1-2a* on chromosome 4 (Jung et al. 2019; Hibrand et al. 2018); Yellow circle with thick black line, *Copia R24588* fragments (226 bp) located within *NUDX1-1a* block #1 on chromosome 2; Yellow circle, corresponding *box38* repeat A; GAG, conserved capsid domain of the retrotransposon polyprotein; LTR, Long Terminal Repeat; ORF, Open Reading Frame. Coordinates are in bp.

Fig. 9. Confocal laser scanning microscopy of transient expression of GFP constructs in agroinfiltrated petals of Old Blush.
a. Schematic maps of constructs including respectively 1085 bp, 521 bp, 316 bp, 138 bp upstream *RcNUDX1-1a,* 1529 bp upstream *RcNUDX1-1b,* and *GFP* alone (empty vector).
b to **l.** Confocal images except for **d**, **f** and **k** taken by reflection of light on the preparation. Petals were infiltrated with the following constructs: *35S:GFP* (**b** and **c**), empty vector (**d** and **e**), *b1529:GFP* (**f** and **g**), *a1085:GFP* (**h**), *a521:GFP* (**i**), *a316:GFP* (**j**), and *a138:GFP* (**k** and **l**).

Cloning was made with FP11-RP11 to FP15-RP11 primers (supplementary table S7, Supplementary Material online). Scale bars, 20 μm.

Fig. 10. Scenario of evolution of *NUDX1* in botanical roses.

a. Global scenario of duplications and specializations. Step 1, specialization of an unknown ancestral *NUDX1* into *NUDX1-3*; Step 2, *cis*-duplication of *NUDX1-3*; Step 3, specialization of *NUDX1-3* into *NUDX1-1b* and *NUDX1-2a* (during this step some TEs were probably inserted near *NUDX1-2a*); Step 4, *trans*-duplications of *NUDX1-1b* and *NUDX1-2a* (after this step, *NUDX1-2a* could have pseudogenized); Step 5, functionalization of expression in petals (during this step *box28* could have duplicate); Step 6, *cis*–duplications of *NUDX1-1a* and increase of the level of geraniol emission.

b. Example of possible *RcNUDX1-1b* to *RcNUDX1-1a* transposition.

Large white arrow, putative ancestral *NUDX1* gene; large orange arrows, genes from Nudx1-1 clade; large blue arrows, genes from Nudx1-2 clade; large green arrows, genes from Nudx1-3 clade; pink drawings, MITE *P580.2030*; dark blue drawings, MITE *G13554*; yellow arrow, *Copia R24588*; dashed grey arrows, specialization steps; black arrows, duplication steps; orange curvy arrows, volatile emission; Chr, chromosome.

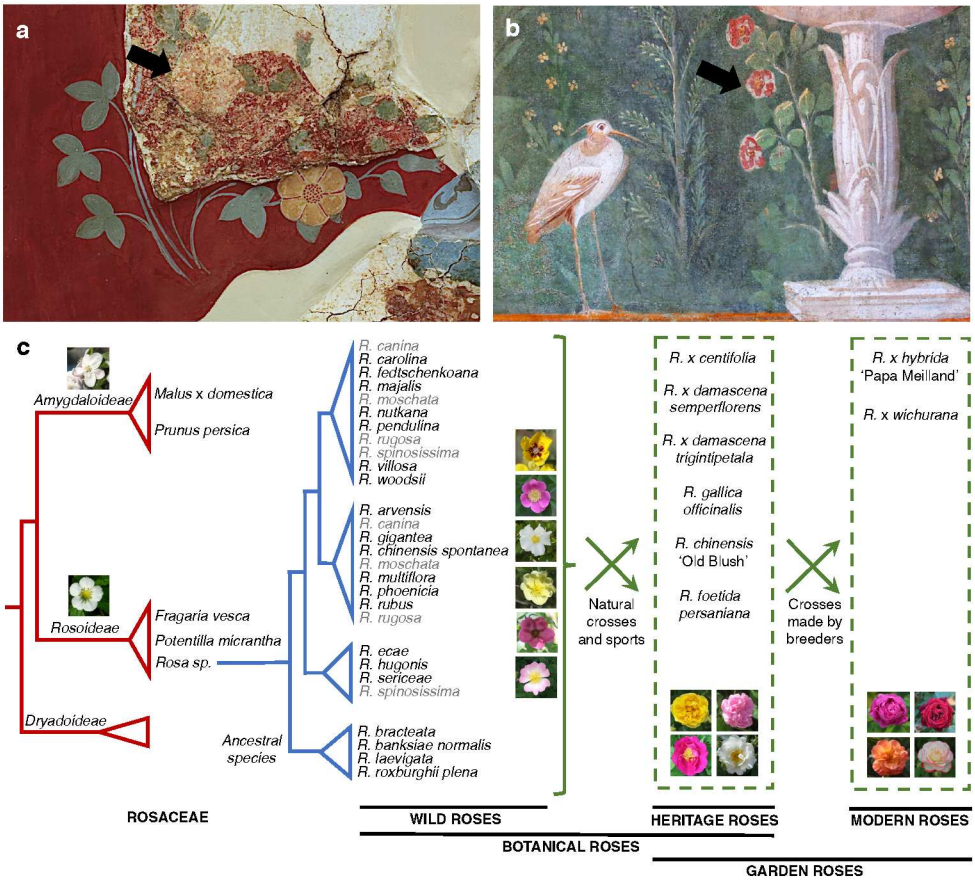
a
b
c
Amygdaloideae
Malus x domestica
Prunus persica
Rosoideae
Fragaria vesca
Potentilla micrantha
Rosa sp.
Dryadoideae
Ancestral species
R. canina
R. carolina
R. fedtschenkoana
R. majalis
R. moschata
R. nutkana
R. pendulina
R. rugosa
R. spinosissima
R. villosa
R. woodsii
R. arvensis
R. canina
R. gigantea
R. chinensis spontanea
R. moschata
R. multiflora
R. phoenicia
R. rubus
R. rugosa
R. ecae
R. hugonis
R. sericeae
R. spinosissima
R. bracteata
R. banksiae normalis
R. laevigata
R. roxburghii plena
Natural crosses and sports
R. x centifolia
R. x damascena semperflorens
R. x damascena trigintipetala
R. gallica officinalis
R. chinensis 'Old Blush'
R. foetida persaniana
Crosses made by breeders
R. x hybrida 'Papa Meilland'
R. x wichurana
ROSACEAE
WILD ROSES
HERITAGE ROSES
MODERN ROSES
BOTANICAL ROSES
GARDEN ROSES

Fig.1

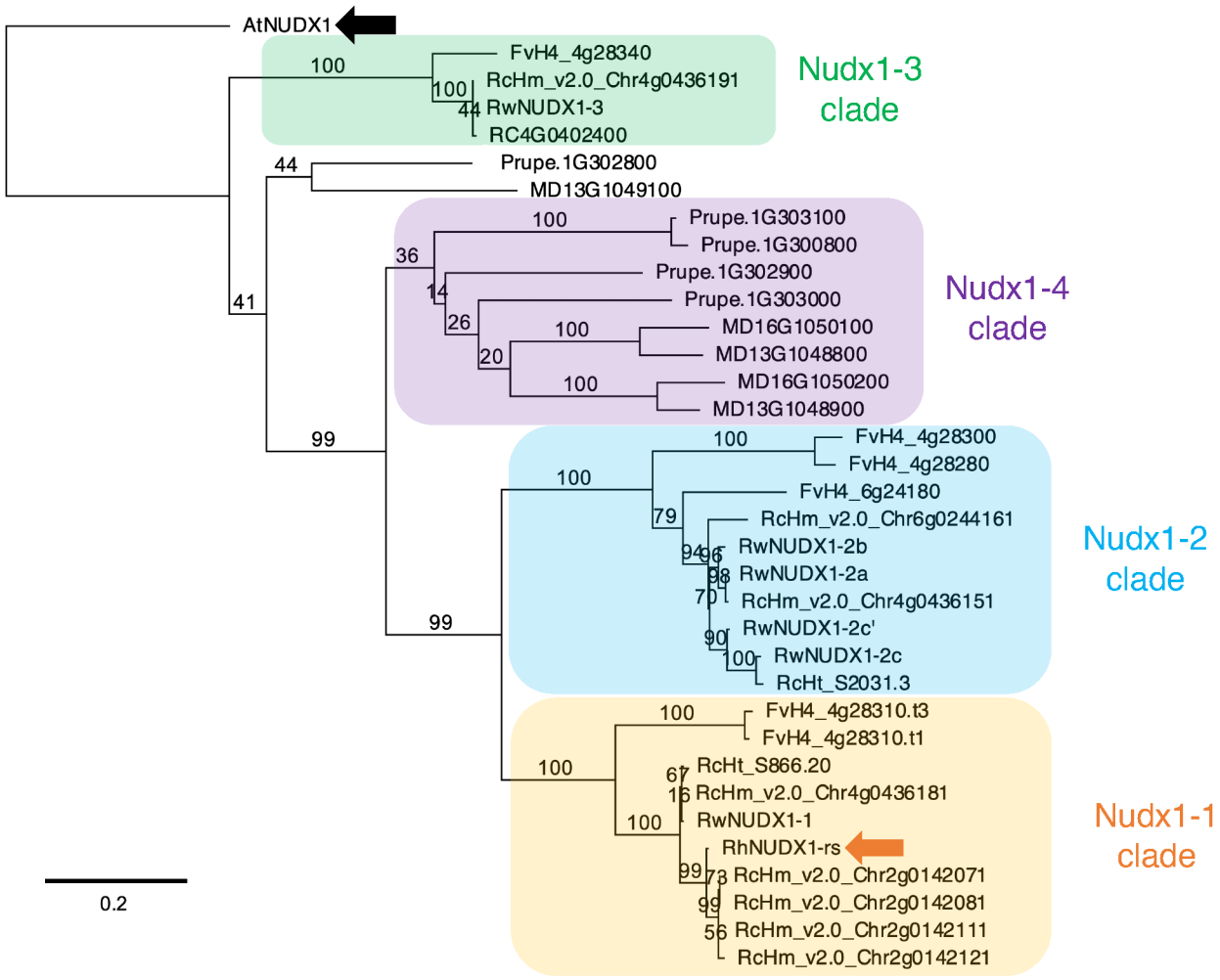
AtNUDX1
FvH4_4g28340
RcHm_v2.0_Chr4g0436191
RwNUDX1-3
RC4G0402400
Nudx1-3 clade
Prupe.1G302800
MD13G1049100
Prupe.1G303100
Prupe.1G300800
Prupe.1G302900
Prupe.1G303000
MD16G1050100
MD13G1048800
MD16G1050200
MD13G1048900
Nudx1-4 clade
FvH4_4g28300
FvH4_4g28280
FvH4_6g24180
RcHm_v2.0_Chr6g0244161
RwNUDX1-2b
RwNUDX1-2a
RcHm_v2.0_Chr4g0436151
RwNUDX1-2c'
RwNUDX1-2c
RcHt_S2031.3
Nudx1-2 clade
FvH4_4g28310.t3
FvH4_4g28310.t1
RcHt_S866.20
RcHm_v2.0_Chr4g0436181
RwNUDX1-1
RhNUDX1-rs
RcHm_v2.0_Chr2g0142071
RcHm_v2.0_Chr2g0142081
RcHm_v2.0_Chr2g0142111
RcHm_v2.0_Chr2g0142121
Nudx1-1 clade
0.2

Fig. 2

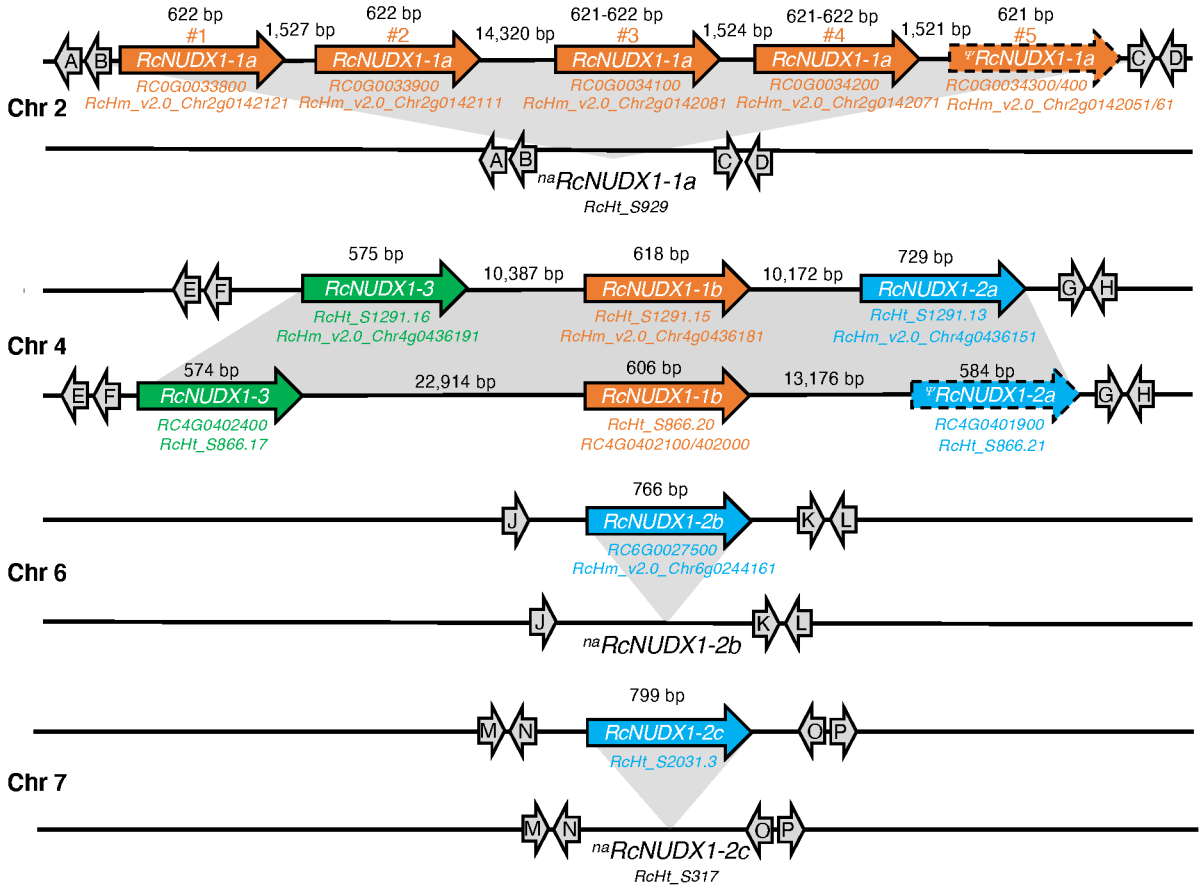
Chr 2
622 bp
#1
RcNUDX1-1a
RC0G0033800
RcHm_v2.0_Chr2g0142121
1,527 bp
622 bp
#2
RcNUDX1-1a
RC0G0033900
RcHm_v2.0_Chr2g0142111
14,320 bp
621-622 bp
#3
RcNUDX1-1a
RC0G0034100
RcHm_v2.0_Chr2g0142081
1,524 bp
621-622 bp
#4
RcNUDX1-1a
RC0G0034200
RcHm_v2.0_Chr2g0142071
1,521 bp
621 bp
#5
ΨRcNUDX1-1a
RC0G0034300/400
RcHm_v2.0_Chr2g0142051/61
A
B
C
D
naRcNUDX1-1a
RcHt_S929
Chr 4
575 bp
RcNUDX1-3
RcHt_S1291.16
RcHm_v2.0_Chr4g0436191
10,387 bp
618 bp
RcNUDX1-1b
RcHt_S1291.15
RcHm_v2.0_Chr4g0436181
10,172 bp
729 bp
RcNUDX1-2a
RcHt_S1291.13
RcHm_v2.0_Chr4g0436151
E
F
G
H
574 bp
RcNUDX1-3
RC4G0402400
RcHt_S866.17
22,914 bp
606 bp
RcNUDX1-1b
RcHt_S866.20
RC4G0402100/402000
13,176 bp
584 bp
ΨRcNUDX1-2a
RC4G0401900
RcHt_S866.21
Chr 6
766 bp
RcNUDX1-2b
RC6G0027500
RcHm_v2.0_Chr6g0244161
J
K
L
naRcNUDX1-2b
Chr 7
799 bp
RcNUDX1-2c
RcHt_S2031.3
M
N
O
P
naRcNUDX1-2c
RcHt_S317

Fig. 3

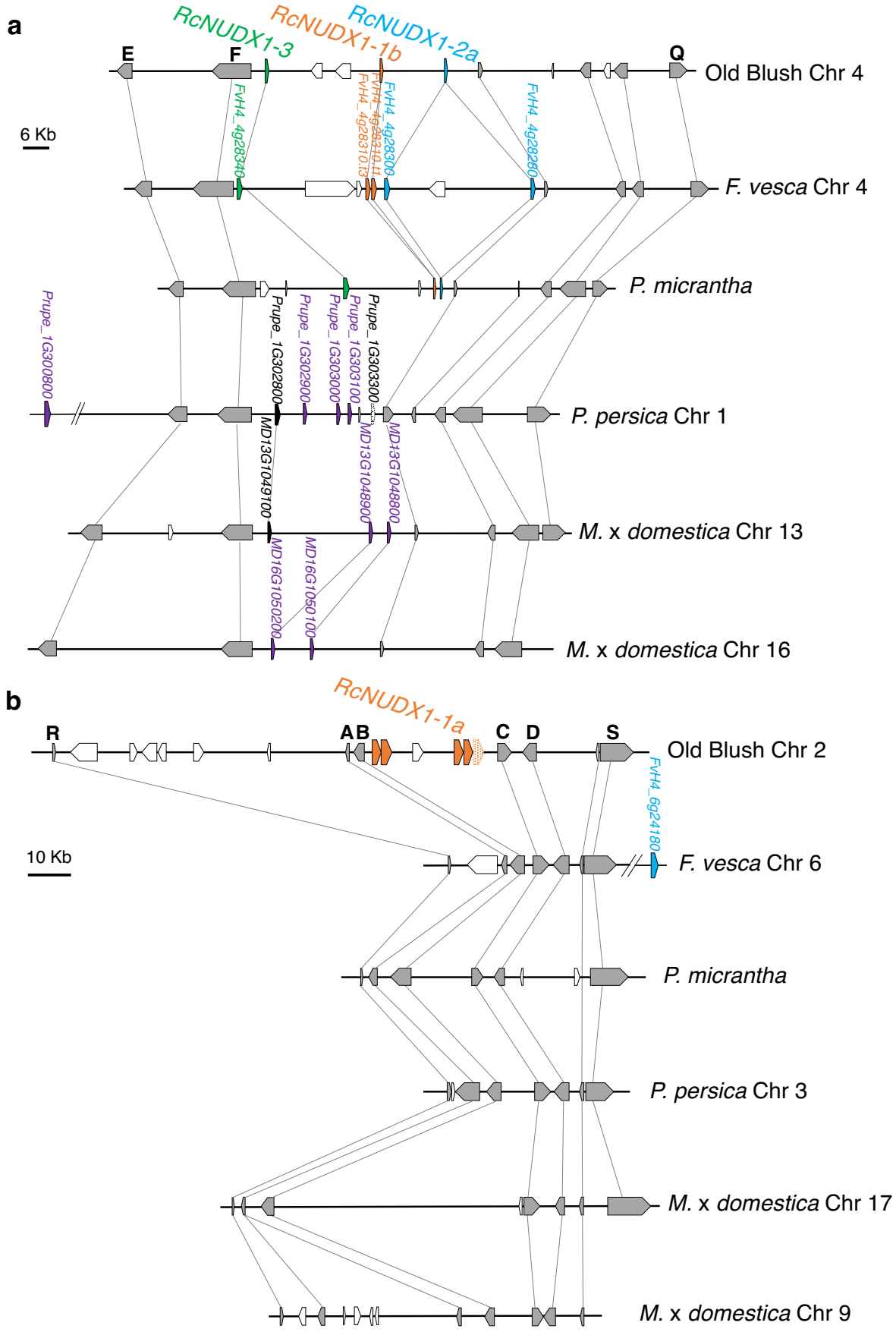
a
RcNUDX1-3
RcNUDX1-1b
RcNUDX1-2a
E
F
Q
Old Blush Chr 4
6 Kb
FvH4_4g28340
FvH4_4g28310.t1
FvH4_4g28310.t3
FvH4_4g28300
FvH4_4g28280
F. vesca Chr 4
P. micrantha
Prupe_1G300800
Prupe_1G302800
Prupe_1G302900
Prupe_1G303000
Prupe_1G303100
Prupe_1G303300
P. persica Chr 1
MD13G1049100
MD13G1048900
MD13G1048800
M. x domestica Chr 13
MD16G1050200
MD16G1050100
M. x domestica Chr 16
b
RcNUDX1-1a
R
AB
C
D
S
Old Blush Chr 2
FvH4_6g24180
10 Kb
F. vesca Chr 6
P. micrantha
P. persica Chr 3
M. x domestica Chr 17
M. x domestica Chr 9

Fig. 4

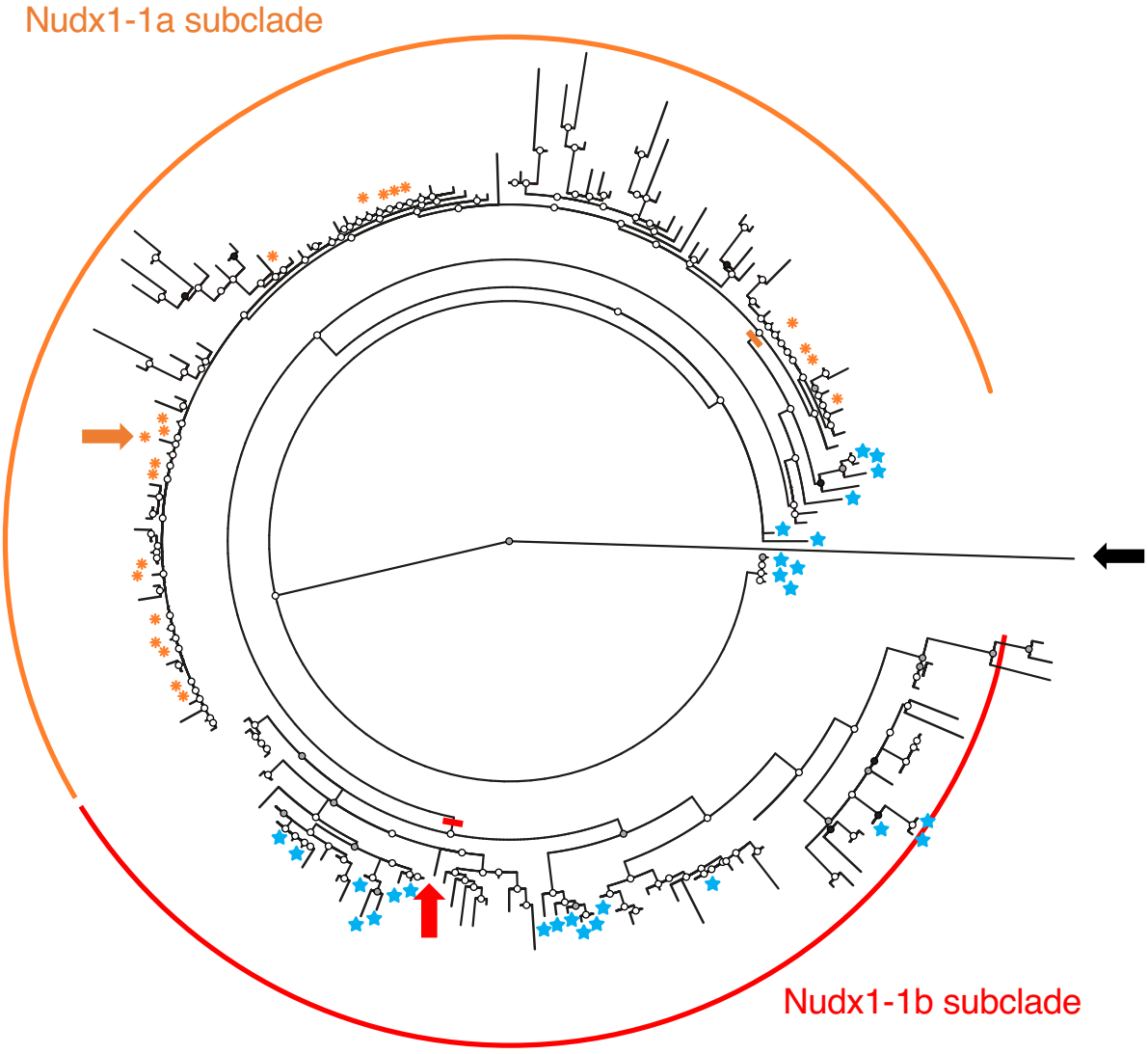

Fig. 5

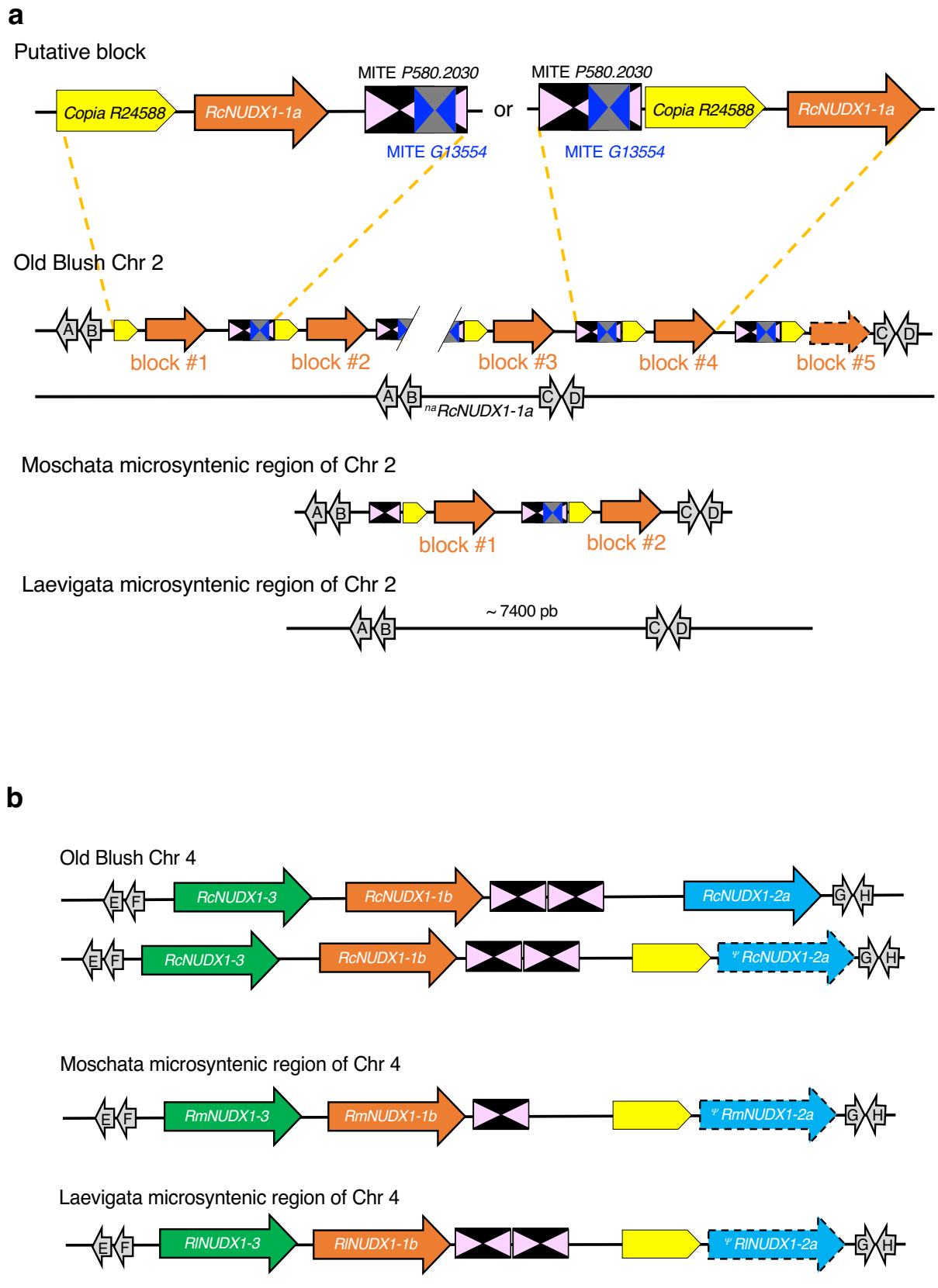
a
Putative block
MITE P580.2030
Copia R24588
RcNUDX1-1a
MITE G13554
or
MITE P580.2030
Copia R24588
RcNUDX1-1a
MITE G13554
Old Blush Chr 2
A
B
block #1
block #2
block #3
block #4
block #5
C
D
A
B
naRcNUDX1-1a
C
D
Moschata microsyntenic region of Chr 2
A
B
block #1
block #2
C
D
Laevigata microsyntenic region of Chr 2
A
B
~ 7400 pb
C
D
b
Old Blush Chr 4
E
F
RcNUDX1-3
RcNUDX1-1b
RcNUDX1-2a
G
H
E
F
RcNUDX1-3
RcNUDX1-1b
ΨRcNUDX1-2a
G
H
Moschata microsyntenic region of Chr 4
E
F
RmNUDX1-3
RmNUDX1-1b
ΨRmNUDX1-2a
G
H
Laevigata microsyntenic region of Chr 4
E
F
RlNUDX1-3
RlNUDX1-1b
ΨRlNUDX1-2a
G
H

Fig. 6

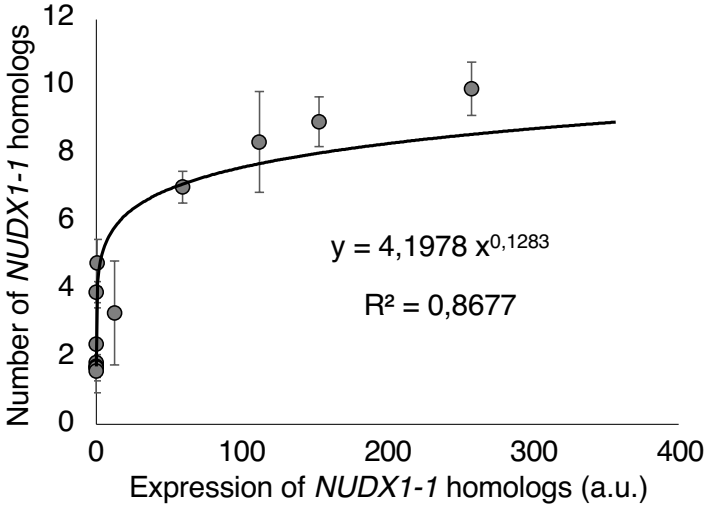

Fig. 7

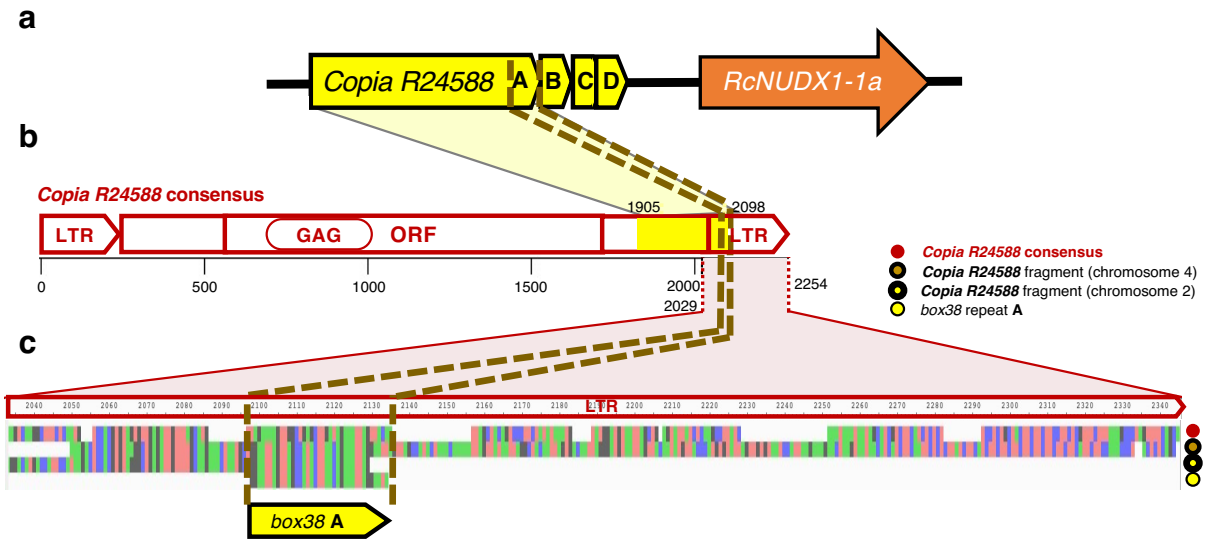

Fig. 8

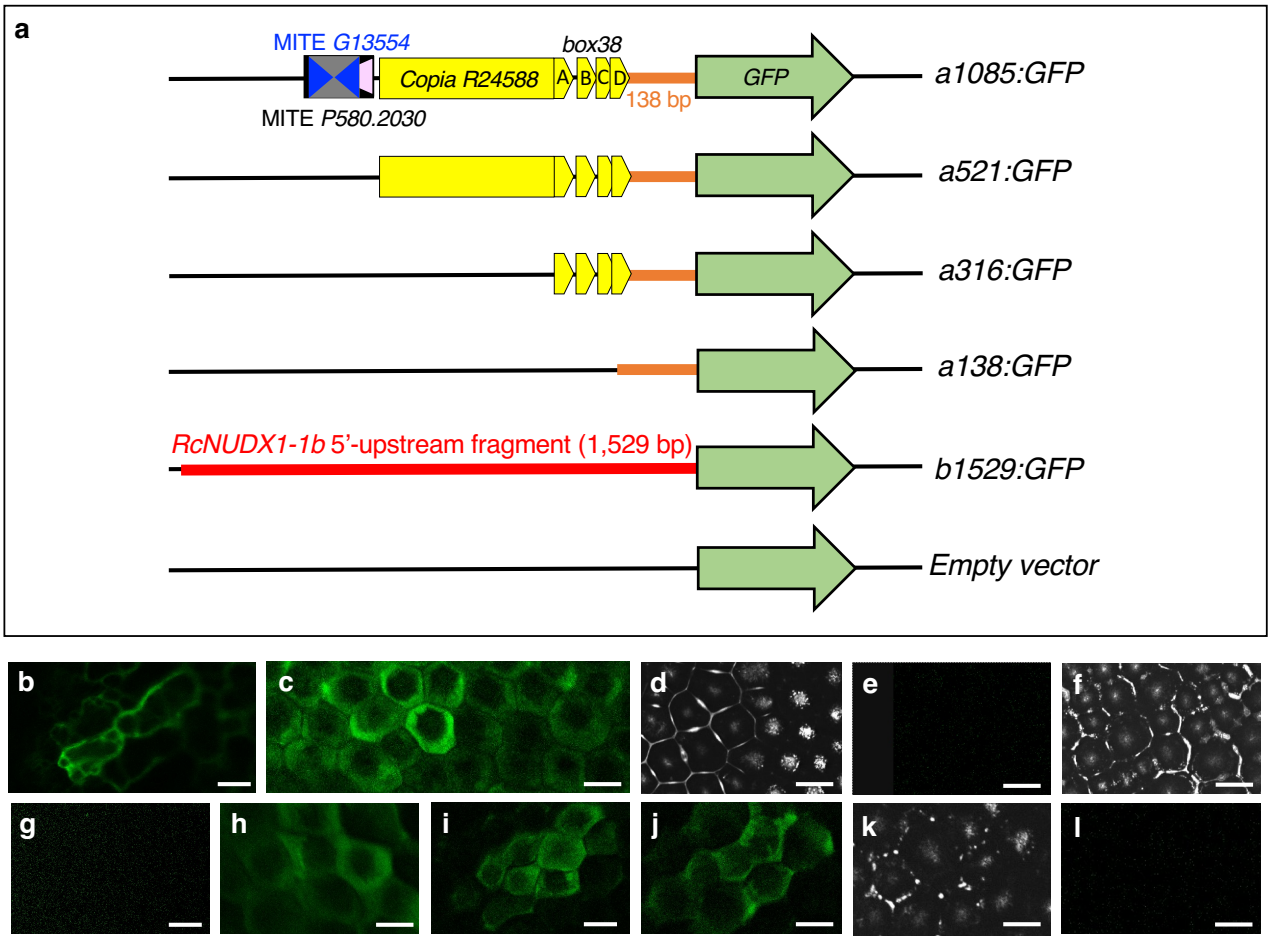

Fig.9

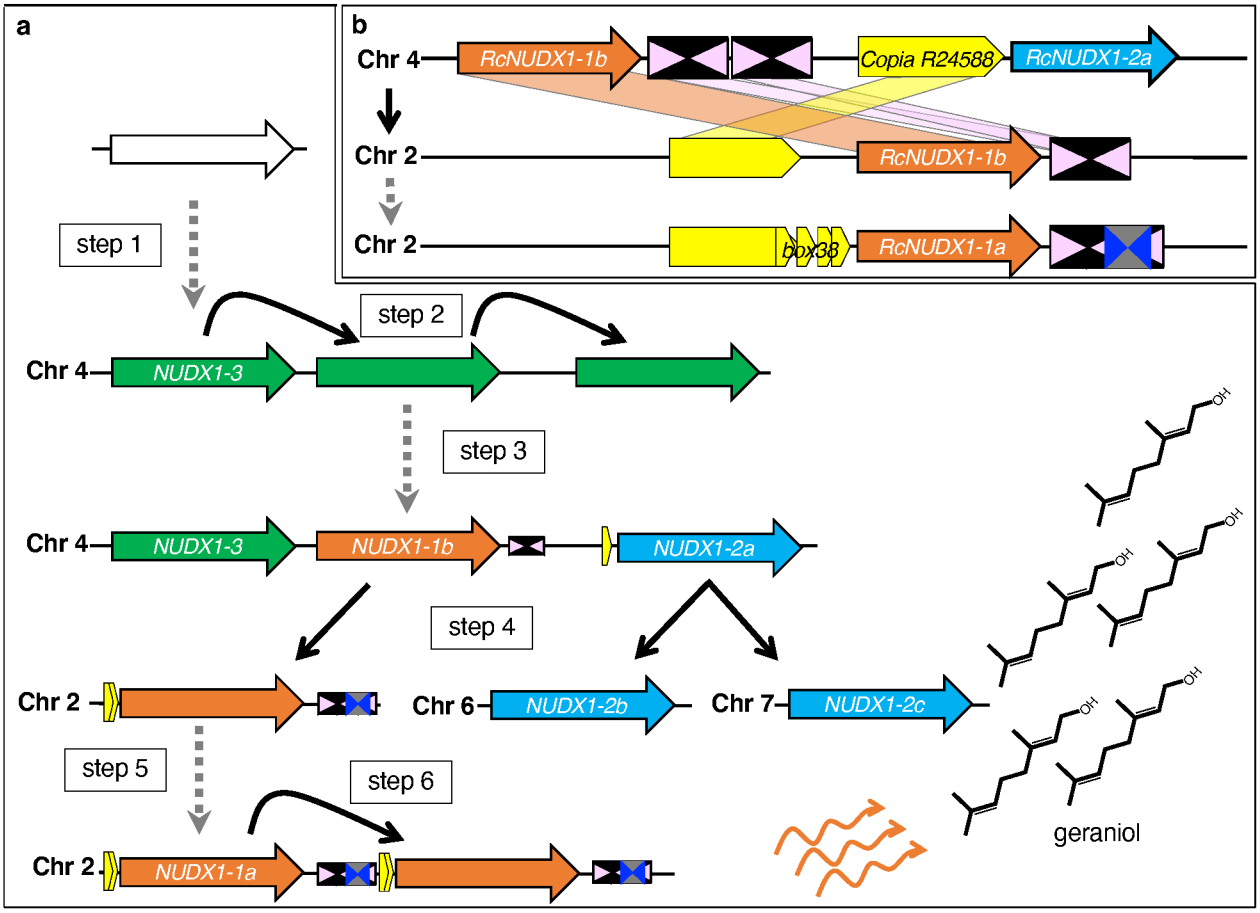
a
b
Chr 4
RcNUDX1-1b
Copia R24588
RcNUDX1-2a
Chr 2
RcNUDX1-1b
Chr 2
box38
RcNUDX1-1a
step 1
step 2
Chr 4
NUDX1-3
step 3
Chr 4
NUDX1-3
NUDX1-1b
NUDX1-2a
step 4
Chr 2
Chr 6
NUDX1-2b
Chr 7
NUDX1-2c
step 5
step 6
Chr 2
NUDX1-1a
OH
geraniol

Fig. 10