Phylogenomics of the genus Rosa : hybridization and polyploidy as factors for diversification

Kévin Debray

To cite this version:

UNIVERSITE ECOLOGIE
BRETAGNE GEOSCIENCES
LOIRE AGRONOMIE ALIMENTATION

université
angers

THÈSE DE DOCTORAT DE

L'UNIVERSITE D'ANGERS
COMUE UNIVERSITÉ BRETAGNE LOIRE

Ecole Doctorale n° 600
Ecole doctorale Ecologie, Géosciences, Agronomie et Alimentation
Spécialités : Écologie et évolution;
Génétique, génomique et bio-informatique

Par

Kevin DEBRAY

Phylogenomics of the genus *Rosa*

Hybridization and polyploidy as factors for diversification

Thèse présentée et soutenue à Agrocampus Ouest centre d'Angers, le 2 mars 2020
Unité de recherche : UMR 1345 IRHS
Thèse N° : (/!\)

Rapporteurs avant soutenance :

Anne Bruneau	Professeure	Université de Montréal, CA
Sophie Nadot	Professeure	Université de Paris-Sud, FR

Composition du Jury :

Président	À définir		
Rapporteurs	Anne Bruneau	Professeure	Université de Montréal, CA
	Sophie Nadot	Professeure	Université de Paris-Sud, FR
Examinateurs	Didier Peltier	Professeur	Université d’Angers, FR
	René Smulders	Professeur	Wageningen University & Research, NL
Directeur de thèse	Fabrice Foucher	Directeur de Recherche	INRA, FR
Co-encadrant de thèse	Valéry Malécot	Maître de conférences	Agrocampus Ouest, FR

Phylogenomics of the genus *Rosa*: Hybridization and polyploidy as factors for diversification

Kevin Debray

March 2, 2020

This dissertation is submitted for the degree of Doctor of Philosophy

University of Angers

Copyrights

Non-plagiarism declaration

I, the undersigned Kevin Debray, declare that I am fully aware that plagiarism of documents or part of a document published on any medium, including the Internet, constitutes copyright infringement and manifest fraud.

Consequently, I commit to cite all the sources I used to write this dissertation.

December 7, 2019

Debray

Je, soussigné Kevin Debray, déclare être pleinement conscient que le plagiat de documents ou d'une partie d'un document publiée sur toutes formes de support, y compris l'internet, constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée.

En conséquence, je m'engage à citer toutes les sources que j'ai utilisées pour écrire ce manuscrit de thèse.

Le 7 décembre 2019

Debray

Abstract

The genus *Rosa* comprises 150-200 species well-distributed throughout the northern hemisphere and presents a complex evolutionary history. Both hybridization and polyploidy represent major driving forces in *Rosa*, yet these two processes were barely investigated in previous phylogenetic studies. With the recent acquisition of whole genome sequencing data and the development of high-throughput sequencing (HTS) techniques, the objective was to develop a general phylogenomics framework to address the phylogenetic relationships in large taxonomic groups made of closely related taxa such in *Rosa*. A mining strategy first identified 1856 informative single-copy orthologous tags ($\mathrm{SCO_{Tag}}$s) in publicly available whole genome sequencing datasets. Ninety-two $\mathrm{SCO_{Tag}}$s from the nuclear genome and four $\mathrm{SCO_{Tag}}$s from the chloroplast genome were then targeted using microfluidic PCRs and amplicon sequencing in a broader sampling of *Rosa* representing 126 species. The HTS data obtained for each accession enabled to estimate ploidy levels and to assemble allelic sequences that further served to trace the origin of hybrid taxa. A stepwise strategy was developed to gradually unveil the reticulated patterns in *Rosa*. Robust plastid and nuclear phylogenies were obtained as well as detailed hybridization scenarios for several specimens. Finally, the resolving power of microsatellite markers was investigated to delineate close species relationships. Using this stepwise framework, many phylogenetic relationships in large and complex taxonomic groups could now be addressed.

Keywords: Phylogenomics; *Rosa* sp.; amplicon sequencing; orthology assignment; networks

Acknowledgments

C'est toujours un sentiment étrange que d'écrire cette section des remerciements, à l'image d'une page qui se tourne sur un grand chapitre de la vie, fait de rencontres et de moments partagés qu'il faut déjà hélas ranger du côté des souvenirs. Le temps passe si vite parfois. L'aventure de cette thèse ne s'est pas faite en solitaire et je voudrais ici remercier ceux qui m'ont accompagné durant ces dernières années.

En premier lieu, je tiens à remercier mes encadrants de thèse, Fabrice Foucher et Valéry Malécot, pour avoir tenu la promesse et l'engagement de me donner les moyens de réaliser cette thèse dans les meilleures conditions. Je mesure bien la chance que j'ai eu de conduire mes recherches dans l'équipe GDO. Merci de m'avoir fait pleine confiance sur ce projet, j'y ai beaucoup appris et j'en ressors grandi.

Merci à Abdelkhader Aïnouche et Alain Franc d'avoir fait partie de mon comité de suivi individuel. J'ai apprécié nos échanges tout au long du projet de thèse.

Merci à Anne Bruneau et Sophie Nadot d'avoir accepté le rôle de rapportrices de ce présent manuscrit, c'est un honneur. Merci également à Didier Peltier et René Smulders d'endosser le rôle d'examinateurs de ce travail. J'aurais plaisir à échanger avec vous sur ces travaux.

Merci aux financeurs publics, Ministère de l'Enseignement Supérieur et de la Recherche et la région Pays de la Loire, d'avoir soutenu ce projet de recherche.

Un grand merci à tous les membres de l'équipe GDO pour l'accueil que vous m'avez réservé. Merci en particulier à Jérémy Clotault pour m'avoir initié à l'enseignement à la faculté des sciences de l'UA et pour nos réflexions intéressantes sur l'évolution des roses. Je remercie également Deepika, Jordan et Subarna pour m'avoir lancé sur les voies de la bioinformatique, j'y ai découvert un véritable trésor. Merci à Tatiana et à Gilles de m'avoir plusieurs fois accompagné faire des prélèvements à la roseraie de Madame Loubert et pour nos émerveillements respectifs sur la diversité de ces plantes. Merci à Annie pour ta bonne humeur et ton aide avec les marqueurs SSRs. Bon courage à la nouvelle recrue, Sophie, que le monde des rosiers ne te soit pas trop épineux. Merci enfin à Éléonore, stagiaire M1, d'avoir rejoint les rangs de l'équipe pour m'assister dans la dernière partie de cette thèse.

Merci aux nombreuses roseraies et jardins botaniques du monde entier qui ont eu la gentillesse de m'envoyer des échantillons d'exception. Sans votre minutieux travail de conservation, rien ne serait sorti de cette thèse. J'ai souvent songé à l'incroyable parcours de ces échantillons, de leur milieu d'origine à mon tube Eppendorf. Merci de m'avoir fait voyager. Merci à Zahra Karimian de m'avoir rapporté d'Iran ces précieux échantillons de *Rosa persica*, cette rose sauvage que j'aime tant.

Merci aux membres de l'équipe EPGV (Marie-Christine Le Paslier, Aurélie Bérard, Élodie Marquand) et de la plateforme de bioinformatique Genotoul (Marie-Stéphane Trotard). Vous faites un travail formidable et j'ai eu beaucoup de chance d'interagir avec vous. Le niveau et la qualité de vos compétences sur ces nouvelles technologies m'ont sincèrement impressionné.

Merci aux doctorants de l'IRHS, Alexis, Élise, Julia, Justine, Mathieu, Marie Anne,... pour le soutien que chacun s'est donné aux cours de nos parcours respectifs, j'ai admiré votre engagement sans limite dans la vie du laboratoire. Je remercie en particulier mes collègues de bureau et de galère, d'abord Marie Anne pour m'avoir montré la voie du doctorat, puis un immense merci à Alexis pour ces moments partagés ensemble dans les joies comme dans l'adversité, d'avoir fait de ce bureau un havre de paix et de douceurs sucrées, j'ai apprécié toutes tes attentions et ta gentillesse. A toi le tour désormais, mais ne t'inquiète pas, ça le fera. Merci aux autres "jeunes" (sans mauvaise blague ;-)) de l'étage, Allan, Séverine, pour les bons moments passés à la cafeteria le midi et ailleurs aussi, je vous souhaite plein de belles choses dans vos carrières respectives.

Veel dank aan mijn Nederlandse leraren in de afgelopen drie jaar: Mercedés, Jana, Ana, Nina en Norbert. Bedankt voor de kennismaking met de Nederlandse en Vlaamse cultuur en voor het aanleren van jullie taal. Deze avondlessen waren een echte verademing in dit zeer wetenschappelijke dagelijkse leven.

Merci aux amis de longue date, Vic, Lolo, Chloé,... pour ces weekends passés ensemble sur Angers et chez vous, de Bordeaux à Paris, sans oublier Pierre-B. On se voit grandir depuis tant d'années et je suis très heureux de vous avoir dans mon entourage depuis si longtemps. Merci à Adrien, Claire, Floflo, Vava, les Dubos pour ces bons moments passés ensemble, ces escapades qui m'ont ressourcées, des bords de Loire à Madrid en passant par Côme.

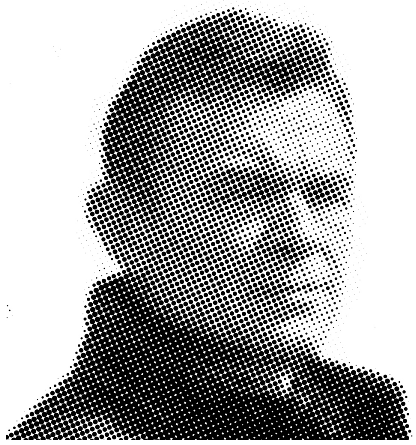
Enfin, merci à mes parents, vous m'avez toujours poussé à chercher le meilleur et à ne pas me contenter de peu. Cette réussite c'est aussi la vôtre. Je suis si fier de toutes ces belles valeurs que vous m'avez transmises et qui me font avancer chaque jour. Merci à la fratrie, Alex et Aymerick, vous êtes courageux, je suis très fier de vous. Merci à ma tante Krystèle pour ton soutien.

“Le Boeing 747 avait à peine décollé de la piste cinquante-neuf que ses 612 passagers eurent déjà à subir des turbulences. A l’avant, Murphy n’en menait pas large, c’était son premier long courrier au-dessus de l’Atlantique et il était impressionné par la rapidité à laquelle l’engin se propulsait : 457 km/h pouvait-il déjà lire sur le compteur digital. Dans très exactement 87 minutes, son co-équipier Jack prendrait le relai et il pourrait alors se détendre en écoutant sa compilation de tubes des années 90s. Pour l’heure il était minuit moins vingt-cinq et l’appareil se stabilisait enfin, quelque part au-dessus des Midlands.”

Aaron Baker, dans *Les aventures de Scarlette Bourgheri au Tisha* (2001).

Dedication

À mon cher Papa.



IN MEMORIAM
1966-2018

Forewords

The present dissertation deals with the evolutionary history of the genus *Rosa* which has been assessed through phylogenomics. I have always been keen on plant diversity, whether natural or induced, since it represents the ground for any breeding programs. Through my studies, I became particularly interested in pre-breeding activities that consist in identifying and transferring desirable traits from a raw, unadapted material to an intermediate set of material that breeders can use further in the development of new varieties. At the end of my Master's Degree, I was torn between directly find a position in a seed company and carrying on my studies to develop new skills. I was particularly concerned by the computational biology associated with the emergence of an ever-growing number of more accurate and affordable sequencing techniques, yet I didn't know anything about bioinformatics. Therefore, I wanted to extend my skills in computational biology, having in mind that the need for such skills would rise in the near future. This is basically how I undertook a PhD project that intended to use phylogenomics to decipher species evolutionary relationships between wild roses.

The PhD project, named PHYROSE, started in October 2016 and finished three years later in October 2019, with an overall funding of € 113,000. The project was supported through a grant from the French Ministry of Higher Education and Research to cover the personnel costs and a grant from the regional program 'Objectif Végétal' (French Pays de la Loire region, Angers Loire Métropole, and the European Regional Development Fund) to fund the running costs and equipment. PHYROSE was conducted at the Research Institute of Horticulture and Seeds (IRHS: Institut de Recherche en Horticulture et Semences, UMR 1345) which gathers the main actors of plant science research in Angers, France (INRA, Agrocampus Ouest, University of Angers). The PhD project was supervised by Valéry Malécot, PhD (Assistant Professor, Agrocampus Ouest) and Fabrice Foucher, PhD (Director of Research, INRA). The Genetic and Diversity of Ornamentals (GDO) team aims to (1) understand the genetic determinism of flowering and resistance traits, (2) investigate the impact of human practices and natural selection on the genetic diversity of ornamental plants, (3) provide requestors with support and knowledge on the above-mentioned aspects. Most of the researches conducted in the GDO team concern the rose. From 2013 to 2016, Mathilde Liorzou and coll. explored the genetic diversity of French rose varieties cultivated throughout the 19$^{\text{th}}$ century, shedding light on a shift from a European to an Asian genetic background. In contrast with cultivated material, the genetic diversity of *Rosa* species was barely investigated in the team, except for few species such as *Rosa gallica*. At the beginning of the PHYROSE project, an international consortium coordinated by the GDO team released a high-quality genome sequence for roses, as well as read sequencing data for eight *Rosa* species. These recent advances provided new opportunities to study the genetic diversity among the wild relatives of cultivated roses. Studying the phylogenetic relationships of *Rosa* at the genome-wide level was challenging due to the relative recent research area of phylogenomics and the inherent complexity of the genus. However, the computational skills that I developed during the thesis have allowed me to broaden the perspectives for the phylogeny of *Rosa*, therefore leading to the most comprehensive evolutionary history of this genus to date.

This dissertation has been written for the degree of Doctor of Philosophy of the University of Angers. The dissertation starts with a first chapter that largely introduces the genus *Rosa* at different levels, from history to genetics, through botany, phylogeny and classification. Relevant methods are also presented. Each of the following chapters has been written in the form of papers for publication. Therefore, some overlapping in writing could not be avoided. The second chapter presents a general method for developing an informative set of phylogenomic markers, considering the genus *Rosa* as an example. This chapter has been accepted for publication in BMC Evolutionary Biology. The third chapter corresponds to the phylogenomic analysis of *Rosa* relationships, considering both hybridization and polyploidy as major driving forces. This article has been submitted. The fourth chapter investigates the extent to which genotyping can resolve close species relationships that were out of the reach of the model-based phylogenomic analyses. The thesis ends with a fifth chapter that corresponds to a general discussion on the overall study.

K.

Contents

List of Figures

List of Supplementary Figures

List of Tables

List of Supplementary Tables

1 A review on wild roses (*Rosa* sp.): History, evolution and classification

1.1 Introduction

Among ornamental plants, the rose undoubtedly holds a privileged place in the heart of people and is found countless times in human beliefs. Sometimes as a symbol of love and sometimes as a symbol of war, the rose fuels the passions of people since the first civilizations. But the rose is actually plural, because it does not refer to one particular species but instead consists of a group of intertwined taxa which embrace a large diversity (Rehder, 1940; Fougère-Danezan et al., 2015). Therefore, we should not speak about the rose in the singular but rather about roses, in reference to the genus *Rosa*. The story of roses began several million years ago (DeVore and Pigg, 2007; Fougère-Danezan et al., 2015), long time before the beginning of mankind. The vast distribution of roses throughout the northern hemisphere can be seen as a great evolutionary success. Evolutionary mechanisms, such as fertile interspecies crosses, may have contributed to rapid adaptations of roses to new environmental conditions during the past millennia (Wissemann and Ritz, 2007). This prefigures a complex evolutionary history where some areas still remain unclear even after decades of studies on the genus. It is also worth mentioning that the strong interest of people in roses have brought even more confusion to the evolutionary history of the genus. Many rose lovers, from botanists to notable amateurs, have indeed grown, described and classified roses, leading to a tremendous amount of names, and the occurrence of hybrid specimens with sometimes unclear wild origins (Brumme et al., 2013; Masure, 2013). The aim of this review is to introduce the originality and the complexity of the genus *Rosa*, which are reflected both in their evolutionary success and in the interest that people have given them. We will first look at the history of roses, from their oldest fossil records to their domestication. Then, we will review the different characteristics that distinguish wild roses from other flowering plants. We will also show how additional characteristics have been used to classify wild roses, from the first botanical censuses to the latest molecular phylogenies. Finally, we will conclude by summarizing the challenges that remain to be met in an attempt to better understand the evolutionary history of these fascinating plants.

1.2 History of wild roses from their oldest fossil records to their domestication

1.2.1 Center of origin and biogeography

Several fossil records found in the northern hemisphere gave insights on the biogeography of wild roses (Table 1) (Becker, 1963; Edelman, 1975; Su et al., 2016). Most of these fossil records only correspond to leaflets, or to entire leaves in the best cases (Figure 1), and many records were given new names while it is difficult to clearly distinguish one record from the other solely based on leaf fragments (Becker, 1963). Therefore, many records may be redundant and could be attributed to one or two morphotypes (Becker, 1963). *R. hilliae* Lesquereux (Figure 1B) and *R. lignitum* Heer (Figure 1C) represent two different morphospecies to which many fossil fragments could be attributed. While *R. lignitum* is commonly found in European paleofloras (Kellner et al., 2012a), *R. hilliae* originate from North America (Becker, 1963). However, they are very close since only sepal persistence might distinguish these two morphospecies (Hably et al., 2000).

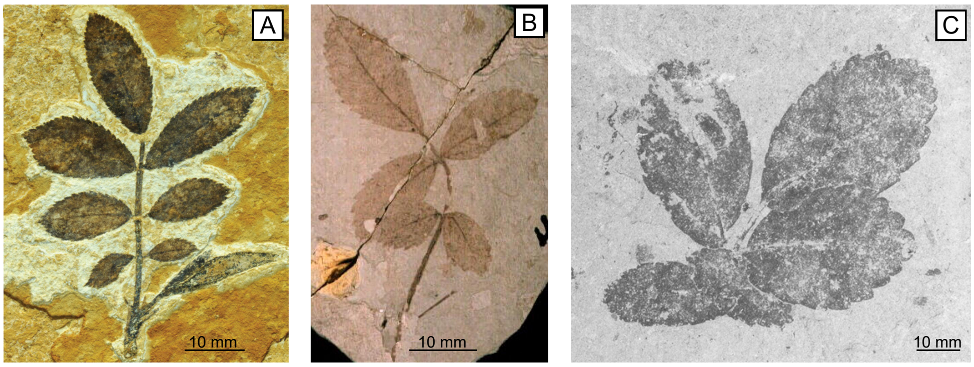


Figure 1: Fossils of wild rose leaf fragments. **A**. *Rosa fortuita* n. sp, Miocene, China (Su et al., 2016). **B**. *Rosa hilliae* Lesquereux (USNM 40575), Eocene, USA (Colorado). **C**. *Rosa lignitum* Heer, Oligocene, from Kellner et al. (2012a) citing Walther and Kvaček (2007)

.

Using time calibrations based on fossil records, Fougère-Danezan et al. (2015) estimated that an early lineage of wild roses evolved during Eocene-Oligocene for 24 MY (54 Mya – 30 Mya), independently from other close related tribes from the subfamily Rosoideae (Figure 2). This early lineage of the genus *Rosa* may have colonized both Asia and North America, since the oldest fossil records attributed to the genus were found in Idaho, USA (Paleo-Eocene) and China (Eocene) (Table 1). The Bering Land Bridge seems to have greatly contributed to genetic exchange between Asian and North American vascular plants (Wen et al., 2016). Transberingian migrations seem to also have occurred in *Rosa* as suggested by the fossil records of Rosa ancestors found in Alaska (Table 1). The climate conditions found during Eocene in these regions was temperate to warm

Table 1: Fossil records of *Rosa*. Adapted from Becker (1963), Edelman (1975), Su et al. (2016) and online database (International Fossil Plant Names Index). Bold species are presented in Figure 1.

Fossil species	Epoch	Region
Rosa cetera Hollick (1936)	Paleo-Eocene	Alaska
R. confirmata Hollick (1936)	Paleo-Eocene	Alaska
R. germerensis Edelman (1975)	Paleo-Eocene	Idaho
R. palaeoacantha Saporta in Heer (1861)	Eocene	France
R. basaltica Ludwig (1858)	Oligocene	Germany
***R. hilliae* Lesquereux (1883)**	Oligocene	Colorado
R. inquirenda Knowlton (1916)	Oligocene	Colorado
R. legányii Andreánsky (1959)	Oligocene	Hungary
R. milosii Kvaček & Walther (2004)	Oligocene	Czech Republic
R. ruskiniana Cockerell (1908)	Oligocene	Colorado
R. scudderi Knowlton (1916)	Oligocene	Colorado
R. sp. Bánhorvati Andreánsky (1959)	Oligocene	Hungary
R. wilmattae Cockerell (1908)	Oligocene	Colorado
R. dubia Weber (1852)	Oligo-Miocene	Germany
R. nausicaes Wess. & Web. (1855)	Oligo-Miocene	Germany
R. sp. var. authors (1961)	Oligo-Miocene	Colorado, Montana
R. sp. var. authors (1962)	Oligo-Miocene	Colorado, Montana
R. angustifolia Ludw. (1860)	Miocene	Germany
R. bohemica Engelhardt (1885)	Miocene	N. Bohemia
R. bursukensis Stephyr. (1987)	Miocene	Moldova
R. chareyrei Boulay (1887)	Miocene	France
***R. fortuita* T. Su et Z.K. Zhou (2015)**	Miocene	S.W. China
R. harneyana Chaney & Ax. (1959)	Miocene	California
R. iljinskiae Stephyr. (1987)	Miocene	Moldova
***R. lignitum* Heer (1869)**	Miocene	E. Europe
R. miocenica Axelrod (1939)	Miocene	California
R. miopannonica Doweld (2017)	Miocene	Austria
R. paraschkevitschii Iljinsk. (1959)	Miocene	Ukraine
R. penelopes Unger (1850)	Miocene	Austria
R. schornii Axelrod (1992)	Miocene	Nevada
R. shanwangensis Hu & Chaney (1940)	Miocene	E. China
R. styriaca Kovar-Eder & Krainer (1988)	Miocene	Austria
R. usuyensis Tanai (1961)	Miocene	Japan
R. akashiensis Miki (1937)	Pliocene	Japan
R. bergaensis Mai & Walther (1988)	Pliocene	Germany
R. Div. spp. Szafer (1947)	Pliocene	Poland
R. hoerneri Chaney (1935)	Pliocene	N.W. China
R. alvordensis Axelrod (1944)	Pliocene	Oregon
R. polyantha Sieb. & Zucc. (sensu Miki, 1937)	Plio-Pleistocene	Japan

temperate which corresponds to current climates where extant species of *Rosa* grow. The *Grande Coupure* marks the transition between the end of Eocene and the beginning of Oligocene and corresponds to a large-scale extinction associated with fauna and flora turnover (Prothero, 1994; Sun et al., 2014) (Figure 2). In Europe, a shift occurred from subtropical-warm temperate, mostly evergreen vegetation in the late Eocene to warm temperate forests that contained almost 40% of deciduous woody plants in Oligocene (Kellner et al., 2012a). During the *Grande Coupure*, the temperature dropped steadily within a very short period (Katz et al., 2008; Liu et al., 2009) and marked the beginning of a new era dominated by an alternation of glacial and interglacial periods (Mudelsee et al., 2014). Many *Rosa* species from the early lineage may have not survived this climate change and disappeared. However, Fougère-Danezan et al. (2015) suggest that species from *R.* subg. *Hesperhodos*, currently found in the southern part of North America, may correspond to relics of the early presence of *Rosa* species in America, prior to the *Grande Coupure*. Few million years after this large-scale extinction, two clades of *Rosa* seem to have diverged from species belonging to the early lineage that survived the *Grande Coupure*. These two clades correspond to the *Rosa* (ex *Cinnamomeae*) clade and the *Synstylae* clade, which are commonly found at the base of the genus in phylogenetic analysis based on molecular data (Bruneau et al., 2007; Fougère-Danezan et al., 2015). The split from the early lineage that resulted in the *Rosa* and the *Synstylae* clades seems to have occurred some 30 MYa (Fougère-Danezan et al., 2015), possibly in Asia, nearby the actual center of diversity of the genus *Rosa* (Figure 2). Thereafter, the *Rosa* clade extended eastward to North America, through the Bering Land Bridge, and started a second colonization of North America by *Rosa* species. As for the *Synstylae* clade, it extended westward to Europe. At that time, the Turgai Sea that had separated Europe from Asia for some million years dwindled drastically due to the rapid decrease of temperature associated with the *Grande Coupure* (Tiffney and Manchester, 2001) that trapped water into glaciers. This resulted in the reunification of Europe and Asia and the closure of the Turgai strait may therefore have facilitated the expansion of the *Synstylae* clade in Europe.

The last 30 MY correspond to the evolution of the *Rosa* and the *Synstylae* clades that resulted in the genus *Rosa* as we know it today (Fougère-Danezan et al., 2015). Geological evolution may have greatly contributed to the diversification of the genus *Rosa*. During Cenozoic, the uplift of the Qinghai-Tibet plateau and the orogeny of the Himalayas changed both climate and topography (Qingsong, 2000; Su et al., 2013; Favre et al., 2015). The collision between the Indo-Australian plate and the Eurasian plate generated many craggy landscapes, with different climate conditions (Su et al., 2016) depending on their orientation, altitude, wind and sun exposure. Topographic changes may have also isolated ancestral population that further derived into new species (Jacques et al., 2014). In addition to geological changes, the alternation of glacial and interglacial periods after the *Grande Coupure* may also have exercised a selective pressure on *Rosa* populations forcing them to innovate in their way of growing and multiplying. This resulted in a diverse range of hardier and therefore more accommodated shrubs regarding the frequent climate changes in late Cenozoic (Gao et al., 2015). The relative young *Caninae* lineage is a good example of a rapid

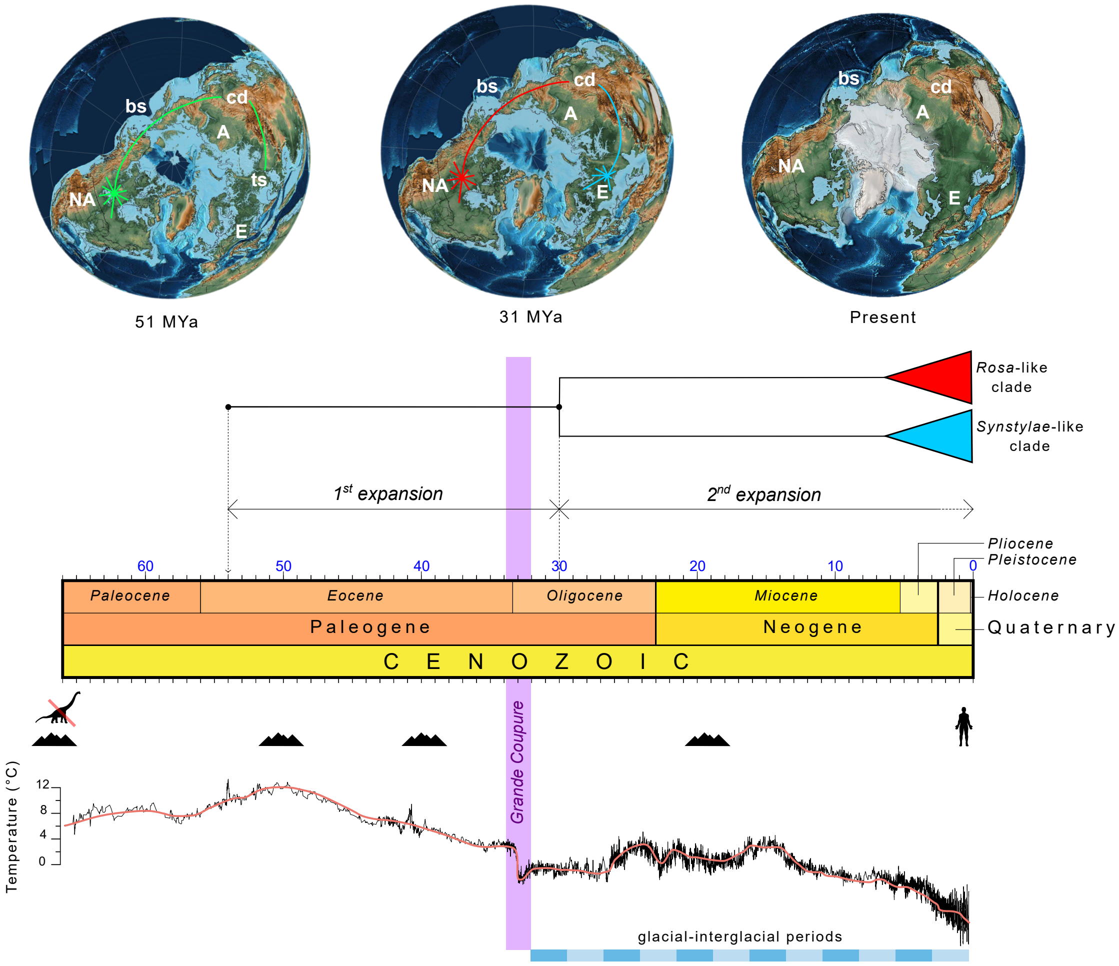


Figure 2: Dynamic evolution of the genus *Rosa* throughout the Cenozoic. The Cenozoic era is marked by the *Grande Coupure* that corresponds to a rapid decrease of temperatures associated with massive flora and fauna extinctions. The genus *Rosa* spread through Northern Hemisphere in two rounds of expansions. The first one occurred before the *Grande Coupure* and corresponds to the expansion of *Rosa* from their center of diversity (cd) in Asia (A), to North America (NA) thanks to the Bering Strait (bs). Expansion westward to Europe (E) was impossible due to the Turgai Strait (ts). The three paleomaps of Northern Hemisphere come from the PaleoAtlas project version 3 and correspond from left to right to times 51 MYa (emergence of the genus *Rosa*, beginning of the first expansion), 31 MYa (end of the first expansion, split in two main lineages after the *Grande Coupure*: *Rosa*-like in red and *Synstylae*-like in blue, and beginning of the second expansion) and Present. Black lines on paleomaps correspond to current states' borders. The Cenozoic time scale is according to the International Stratigraphy Chart and is scaled in Million years ago (MYa). Orogeny events in Northern Hemisphere are represented by mountain shapes. Climate change is represented by the evolution of temperature as estimated from deep-sea oxygen records (extracted from Mudelsee et al. (2014)). The blue bar represents the alternation of glacial-interglacial periods in late Cenozoic but is not time-scaled.

radiation within the genus *Rosa* since it appeared only 6-8 MYa and still represent 20% of the extant species of *Rosa* (Wissemann, 2002; Ritz et al., 2005; Fougère-Danezan et al., 2015). The wide range of habitats generated after orogeny associated with the periodic climate changes may have provided the opportunity for *Rosa* to adapt to these new surrounding conditions and therefore contributed to the diversity currently observed in the genus.

1.2.2 Current distribution of wild roses

Nowadays, wild roses find their center of diversity in the region that forms Central Asia (Henker, 2000). About half of extant species of *Rosa* thrive in this region with 65 species that are endemic to China (Cuizhi and Robertson, 2003). *Rosa* species are found in most of the temperate to subtropical regions of the Northern Hemisphere (Rehder, 1940). This wide distribution encompasses very different habitats that are generally particular to a group of species. For instance, *R. persica* is endemic to arid regions of the Middle East (Iran, Afghanistan, Iraq, Kazakhstan), on salty and stony soils nearby the Caspian and Aral seas, in arid steps of Siberia and West China (Basaki et al., 2009). *R. abyssinica* only thrive in the Horn of Africa (Somalia, Ethiopia, SW of the Arabian Peninsula) and is the southernmost species of the genus *Rosa* (Thulin, 1993; Bein et al., 1996). On the contrary, *R. acicularis* is the northernmost *Rosa* species and is found at high latitudes in Europe, Asia and North America, nearby the Arctic circle (Lewis, 1959). Several wild rose species, generally corresponding to hardy shrubs, displays populations that can be found in very different habitats. This is the case for populations of *R. spinosissima* (syn. *R. pimpinellifolia*) which can be encountered at very different altitudes, both along the British coasts and in the mountainous regions of Central Asia (Cuizhi and Robertson, 2003; Boyd, 2015). Populations of *R. mollis* can also be found at different latitudes in Europe, from Portugal to Finland (GBIF, 2019).

Several wild rose species have been introduced by people into areas of similar longitudes as their region of origin, either intentionally or accidentally. For instance, *R. majalis* is endemic to North and Central Europe but it has naturalized in some areas of NE America (Lewis et al., 2014), adjacent to Canada. *R. montezumae* is a *Caninae*-like species that is found in Mexico while it probably originates from Europe, as other *Caninae* species (Masure, 2013). Some of the introduced rose species are now regarded as invasive species. This is the case for *R. multiflora* that originated from Asia and was introduced in East America for erosion control and as a living fence (Amrine, 2002). *R. multiflora* is now considered an invasive species since it forms dense thickets that invade pastures and crowd out native species (Amrine, 2002). *R. laevigata*, originating from subtropical SE Asian regions, was introduced in US during early colonial times and was further propagated by Cherokee Indians (Tabor, 1960). They distributed *R. laevigata* to such an extent that it has been given the name of ‘Cherokee rose’ and became the state flower of Georgia in 1916 despite the fact that it does not originate from SE USA (Tabor, 1960). Nowadays, *R. laevigata* spreads somewhat aggressively from North Carolina to Florida west to Texas (Miller et al., 2004).

1.2.3 Relationships between roses and people through History

People have always esteemed roses highly, at different times and in many civilizations. Roses are generally associated with the most prosperous periods of history, and the rose has been chosen as a flattering symbol in many different fields, from religion to politics through literature and economics (Foxton-Smythe, 2013).

Ancient History (B.C. 60,000 - 650 A.D.)

Archaeological evidence of roses throughout human history is scarce (Widrlechner, 1981). Excavations in The Netherlands revealed accumulation of rose seeds along with other fruits and nuts remains of *Corylus*, *Pyrus*, *Crataegus*, and *Rubus* at an inhabited site during Neolithic (5000 B.P.) (van Zeist and Palfenier-Vegter, 1981). The authors suggest that humans may have intentionally collected rose hips for food. Small rose seed heaps were also described in German sites (7000 B.P.) (Elburg, 2010) and in Britain Brown and Murphy (2000), hinting at the possibility that rose hips were an integral part of Neolithic man's diet.

The oldest known evidence showing a representation of rose dates back to the time of the Minoan people (3500 B.P.) from the archaic period of the Greek civilization (Widrlechner, 1981). In 1900, Sir Arthur Evans excavated several fragments of what could have been a larger fresco from the Palace of Minos. Rose drawings are present on one of the fragments along with a bird, an iris, a lily and another plant (Tucker, 2004). The fresco was named "The Blue-bird fresco" after the color of the bird and a detail from a modern reconstitution is presented in Figure 3A. It is worth mentioning that the fresco has been restored by 20th century artists that may have introduced biases. On this subject, Tucker (2004) wrote: "when we look at the original Blue Bird Fresco [...] we find that the one original rose [...] is painted slightly twisted to profile, with five overlapping petals that are now a faded pale pink" which greatly contrasts with the modern reproduction of Émile Gilliéron that displays a six-petaled yellow rose (Figure 3A). While Hurst (1941) argues that the rose on the fresco corresponds to *R.* × *richardii*, a hybrid between *R. gallica* L. and *R. phoenicia* Boiss, Tucker (2004) suggests that it may be *R. pulverulenta* (syn. *R. glutinosa*) based on current distribution and the 3-leaflet leaves on flowering shoots. In any case, both species occurred in Greece at this time and present scent-related traits, with pine-scented leaves for *R. glutinosa* and fragrant flowers for *R.* × *richardii.* If species assignments are correct, these representations of scented roses may suggest a possible use of these kinds of roses for their fragrance by the Minoan civilization.

While there is still question regarding whether rose domestication began in China or Mesopotamia (Bombarely, 2018), it is likely that roses were largely domesticated for their scent and the resulting rose water (Widrlechner, 1981). Chinese philosopher Confucius (551-479 B.C.) reported that the imperial library contained hundreds of books on roses (Foxton-Smythe, 2013; Chwalkowski, 2016). The production of rose water was already understood at that time but were kept for privileged persons (Foxton-Smythe, 2013). All the results from hundreds of years of rose breeding in China were only known in Europe from the 18th century with the introduction of the tea-roses (see later) (Guoliang, 2003). Aside from scent purposes, wild roses and their hybrids were associated with rituals in many ancient civilizations. For instance, the Egyptian crafted wreaths made of buds from *R.* × *richardii* for their funeral ceremonies, and very well conserved remains were found in the cemetery of Hawara, Fayum province (Newberry, 1889). Other ancient civilizations associated roses with a special power, like the Chinese who encapsulated dried rose flowers in small pouch that they wore to drive evil spirits away (Foxton-Smythe, 2013). Roses were therefore cultivated a lot by Mediterranean and Chinese civilizations. Greek philosopher and botanist Theophrastus (B. C. 371 – 288) reported on rose diversity, mentioning that double-flower types already exist by that time (Hort, 1916). Even then, the rose was considered as the queen of flowers in the Greek civilization according to statements attributed to the poetess Sappho (Potter, 2011). Later, the Romans inherited the passion for roses from the Greeks and popularized their use, not only for attar production and consumption but also by attributing a wide range of purposes and symbols to roses (Widrlechner, 1981). Rose blossoms started to be used to adorn homes and special powers were given to rose petals such as preserving a woman's youth and beauty (Foxton-Smythe, 2013). Roses were cultivated a lot during the Roman Empire, according to numerous authors including Pliny the Elder (23 – 29 A.D.). The requirement on rose petals was so considerable that there were not enough lands in Italy to grow roses and the Roman had to settle new rose plantations in Egypt (Widrlechner, 1981; Foxton-Smythe, 2013). On this point, Horace complained about the fact that grain fields were gradually replaced by ornamental plantations for the Roman aristocracy which was quite detrimental for the local populations (Conington, 1872). The Romans can be seen as trailblazers in rose breeding and cultivation since they invented the 'winter' rose (*rosae hibernae*), especially adapted to the Egyptian weather conditions, to extend the production of rose earlier in the season (Foxton-Smythe, 2013). It has been reported that Roman emperor Nero (37 – 68 A.D.) paid one ton of gold for one shipment of rose petals from Egypt to Rome (Widrlechner, 1981). Indeed, mass-quantity of roses were necessary to supply Romans feasts where confetti made of rose petals was quite common. *The Roses of Heliogabalus* is a 19th century painting by Sir Lawrence Alma-Tadema that depicts the use of rose petals as confetti during a banquet organized by the Roman emperor Elagabalus (203 – 222 A.D.) (Figure 3C). Tremendous quantities of rose petals fell out of a reversible ceiling, covering the banquet table and the guests, some of whom even died suffocating under the tons of dropped petals (Lampridius, 0 AD). Towards the end of the Roman Empire, there were less wealthy citizens and the need for the luxury rose petals became scarce. Rose plantations were gradually abandoned and only the hardiest roses could return to the wild. However, all the symbols and folklore associated with roses and developed by the ancient civilizations, especially the Roman, persisted long after their decline and were transmitted to the following societies.

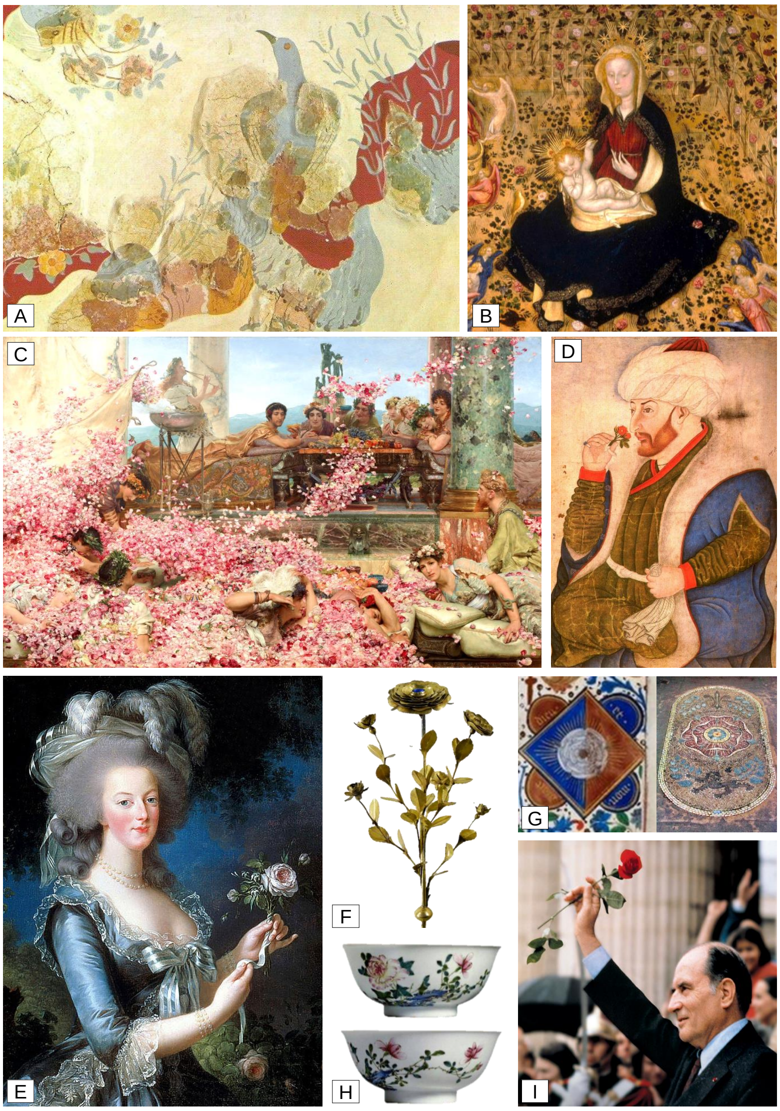


Figure 3: Roses throughout History and culture. **A**. *The blue bird fresco* (detail from a modern reproduction by Émile Guilliéron fils) from the House of Frescoes at Knossos, Late Bronze Age (ca. 1500 B.C.E.), Paintings on wet lime plaster, H. 60cm, Heraklion Museum of Crete, Greece. **B**. Michelino da Besozzo or Stefano da Verona, *Madonna del Roseto* (detail), ca. 1420-1435, Tempera on panel, H.130cm W.95cm, Castelvecchio Museum of Verona, Italy. **C**. *The Roses of Heliogabalus*, Sir Lawrence Alma-Tadema, 1888, Oil on canvas, H.132cm W.213cm, Private collection. **D**. A portrait of Sultan Mehmed II Fatih (The Conqueror) smelling a rose, Nakkaş Sinan Bey, ca 1480, Watercolor on paper, H.39cm W. 27cm, Topkapi Palace Museum, Istanbul, Turkey. **E**. *Marie-Antoinette dit "à la rose"*, Élisabeth Vigée Le Brun, 1783, Oil on canvas, H.113cm W.87cm, Palace of Versailles, France. **F**. Minucchio da Siena, Golden Rose from the Treasury of the Basel Cathedral (detail), 14th century, Gold and glass jewelry, H.60cm, Museum of Cluny, France. **G**. Left: Vincent of Beauvais, translated by Jean de Vignay, *Le miroir historial*, books 1-5, The white Rose of House York from a manuscript of Edward IV of England (f. 3r, vol. 1) (detail), ca. 1480, Parchment codex, H.470mm W.340mm, British Library, U.K.. Right: The Red Rose of House Lancaster, Photography of the cobblestone mosaic from the Williamson Park, Lancaster, U.K., ©Lupin. **H**. Porcelain from the Chinese Qing Dinasty (1723-1735), China. **I**. Inaugural speech of French socialist President Mitterrand at the Pantheon in 1981. Photography. Credits unknown.

Middle Age (500 - 1500)

At the beginning of the Middle Age, while Christianity raised in Europe, the rose was no longer a popular symbol since it was rather seen as a remnant of the lavish excesses of the Roman Empire that had oppressed those who preached Christianity (Touw, 1982). Although early church leaders advised not to plant roses, their warning were ignored and roses gradually became a symbol in both Christian culture and literature (Joret, 1892). Appreciation of roses was heightened in the early second millennia, when crusaders came back from the Middle East with roses and new knowledge about their cultivation and uses (Touw, 1982). It must be said that roses were extremely important in Islam, especially in literature where numerous tales and poems referred to them (Zwemer and Zwemer, 1941). Among Turks and Arabs, the white rose owes its origin to the perspiration of the Prophet. According to Persian poetry, roses were born after that the flowers complained to Allah against the Lotus, which was at that time the queen of all flowers, that she slept all night. In response, Allah created the white rose as the new queen and provided her with thorns so that she would be protected (Joret, 1892; Zwemer and Zwemer, 1941). The red rose appeared after the nightingale fell under the spell of the white rose and flew straight for it but was pierced by the thorns and its blood turned the white rose into a red rose (Joret, 1892; Zwemer and Zwemer, 1941). By the time of crusades, the Persians had mastered the art of rose cultivation for centuries, mainly for the production of rose water and later rose attar, on which they developed highly technical skills. While the Romans only soak rose petals in water, the Persians distilled rose water by boiling rose petals with water therefore obtaining a purer product (Widrlechner, 1981). It is also mentioned that rose oil was made by soaking rose petals with almond, or olive oil to transfer and stabilize the flower scent. The rose industry spread from Persia to the surrounding countries, especially Arabia. The Arabs mastered the art of distillation and developed numerous rose flavoring goods, from fragrances to beverages and desserts. Distilled rose water became extensively traded by the Arabs and was introduced to Europe as well as fragrant rose varieties (Widrlechner, 1981).

In Christian Europe, only the monks and nuns had the leisure and the sensibility to cultivate roses (Joret, 1892). The concept of gardens became more and more prominent at that time and was later associated with the medical benefit of plants. In this field also, numerous properties to cure pains and diseases were given to roses. The medicinal species roughly corresponded to *R. gallica* and *R. canina* and were used to allay fever, inflammation or pain and to stop any excessive flow such as hemorrhage or diarrhea (Touw, 1982). By the end of the Middle Age, one of the first Encyclopedia dedicated to pharmacopeia, the *Hortus Sanitatis* compiled in 1491 by Jacob Meydenbach, devoted no less than four pages to roses, in clear contrast with the few lines or so given to most other plants (Touw, 1982). *Rosa canina*, the dog rose (FR: Rosier des chiens, NL: Hondsroos; DE: Hunds-rose) may owe its name to a very ancient belief that gave its roots special properties for curating rabies, mainly transmitted by dog bites at that time. Asides from the medical properties, many symbols were associated with roses in the medieval Christian societies. A golden rose was usually bestowed to notify papal approbation. For instance, in 1096 Pope Urban

II gave a golden rose to the Count of Anjou to express its gratitude for the Count's devotion and loyalty to the catholic church (Cornides, 1967) (Figure 3F). In Christian beliefs, roses often referred to Virgin Mary, as a symbol of purity and love (Figure 3B). It should be mentioned that roses were already associated with female divinities in Greek and Roman polytheistic civilizations (Aphrodite, Venus), for similar symbols. The Holy Rosary is a catholic prayer that has been dedicated to the worship of Mary and Jesus, and its name refers to the crown of roses that was usually depicted on Mary's representations. In England, the rose became an important heraldic flower especially after the Wars of the Roses (1455-1485) that brought for decades two rival branches of the royal House of Plantagenet into conflict: The House of Lancaster and the House of York. The House of Lancaster was associated with a red rose (Figure 3G), probably *R. gallica* (Le Rougetel, 1988), possibly deriving from escape specimens inherited from the lavish past of the Roman society. The House of York was sporting a white rose (Figure 3G), possibly *R.* × *alba* (Le Rougetel, 1988). In 1485, Henri Tudor, from Lancaster, acceded to the throne of England and married Elisabeth of York thus stopping the wars. They took a bicolor white and red rose variety, possibly of *Rosa damascena*, as the new emblem of the crown, hence reconciling the two Houses (Le Rougetel, 1988). The Tudor rose is still present on the coats of arms of England.

During the Middle Age, the rose as a symbolic emblem became quite common in people's habits and customs and was often used in many different fields such as politics, religion, trade, medicine, literature and paintings. This passion for roses and its associated connotations were culturally transmitted through modern history.

Modern History (1500 - present)

While in the Middle Ages rose bush cultivation was essentially the interest of monasteries, many rose lovers started to set their sights on these plants during the early modern history. The outlets of rose cultivation in medicine were gradually outshined by the ornamental interest of these plants. In this way, some rose species were introduced to increase the phenotypic diversity of roses that were cultivated at this time. For instance, *Rosa foetida*, originating from Anatolia, seems to have been introduced during the 16th century (Joyaux, 2015). At that time, it was the only yellow-flowered rose that people knew. The variety *R. foetida* 'bicolor', which petals are orange red on the top face and yellow on the other side, was already known in the 16th century (Joyaux, 2015). *R. foetida* and its varieties are directly or indirectly behind most of the yellow/orange rose cultivars that we know today (Joyaux, 2015). However, it is worth noting that the cultivation of roses for ornament in the early modern period (17th-18th centuries) was limited in Europe, to the benefit of other ornamental species (*Hyacinthus* sp., *Lilium* sp., *Ranunculus* sp.). Nevertheless, the cultivation of roses for fragrance was still a major activity at that time. The production of true attar of rose, that is essential oil from rose flowers, may have been mastered around 1600 in Persia (Widrlechner, 1981). In his account about Sultan Jehangir's wedding in 1643, Mohammed Achem reported that the canals in the palace gardens were filled with rose petals and water for the occasion. It was a warm day and the queen noticed that a thin, oily and highly aromatic film arose at the surface of the water and she attributed this observation to the effect of the sun and heat on rose flowers. She ordered to collect the oily supernatant and this is how attar of rose may have been discovered. Although a cooling would have been necessary to finally separate essential oil from hot water, Mohammed Achem's account may indicate that extraction of rose essential oil was understood in his time.

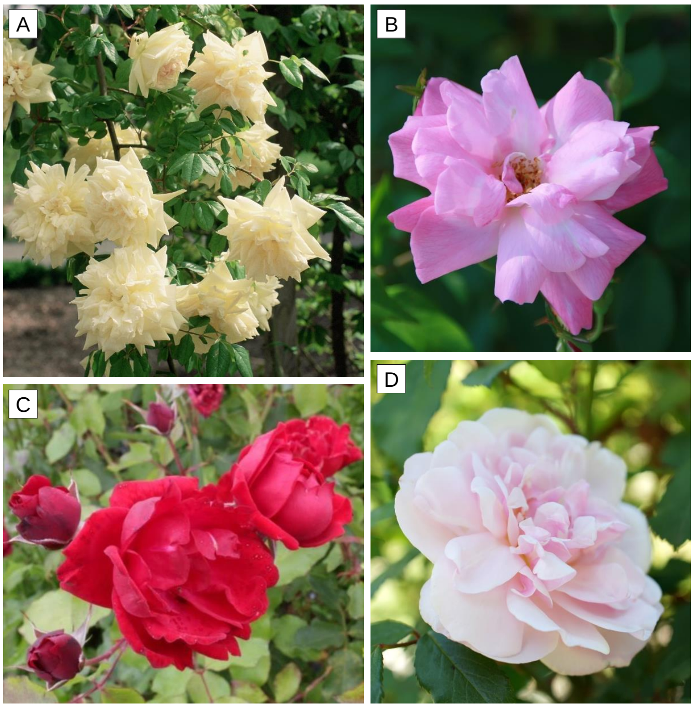


Figure 4: The 'four stud roses'. **A**. 'Park's Yellow Tea-scented China' (Credits: A. Barra). **B**. 'Parson's Pink China' [= *Rosa chinensis* 'Old Blush'] (Credits: David J. Stang). **C**. 'Slater's Crimson China' (Credits: Tasman Bay Roses). **D**. 'Hume's Blush Tea-scented China' (Credits: T. Kiya).

Most of the roses cultivated in Europe or the Middle East were supposed of the European-Mediterranean type. A major landmark in rose breeding history corresponds to the introduction of Chinese cultivars in Europe during the 18$^{\text{th}}$ century (Joyaux, 2015). The history of rose breeding in China was at least as old as the history of rose breeding in Europe/Middle Eastern but was barely known to European rose breeders before. Stories about the introduction of Chinese roses in Europe are very controversial, and the following chronology corresponds to the one presented by Joyaux (2015). The first Chinese rose introduced in Europe may be *Rosa chinensis*, so named by Nikolaus Joseph von Jacquin (1727-1817), based on a specimen that was grown since 1733 by the Dutch botanist Gronovius under the name "Chineeshe Eglantier Roosen" (Joyaux, 2015). The above mentioned *Rosa chinensis* may have been brought back to Europe by Dutch people through the Dutch East India Company. The specimen may actually be *Rosa chinensis* var. *spontanea* (Rehd. & Wils.) Yu & Ku and correspond to an ancient cultivar that has been bred by Chinese people for centuries. In late 18$^{\text{th}}$, another Chinese variety was introduced in England, probably originating from the famous Fa Ti nurseries in Canton. This specimen is called 'Yue Yue Fen' ("monthly pink"), further renamed *Rosa* 'Old Blush' or 'Parson's Pink China' (Figure 4B) after the name of a gardener who cultivated it at Rickmansworth in 1793 when Colville nurseries started to propagate the specimen. Concurrently, a British East India Company agent noticed a red cultivar in Calcutta in 1789, named 'Yue Yue Hong' by the Chinese people and further called *Rosa chinensis* var. *semperflorens* Koehne in Europe. He brought the specimen back to England and gave it to Gilbert Slater, hence it has been known under the name 'Slater's Crimson China' (Figure 4C). Two additional varieties were later introduced by the English in Europe, namely 'Hume's Blush Tea-scented China' (Figure 4D) and 'Parks' Yellow Tea-scented China' (Figure 4A) both supposed to originate from Fa Ti nurseries nearby Canton. 'Hume's Blush Tea-scented China' was supposed to be a hybrid between *Rosa chinensis* var. *semperflorens* and the Asian species *Rosa gigantea*. As for 'Parks' Yellow Tea-scented China', it was appreciated for its yellow color that was quite uncommon in rose gardens at this time. By introducing novelties in color, scent and recurrent blooming, the four Chinese varieties derived from centuries of breeding in China considerably altered the genetic background of the roses that had been cultivated in Europe until then and which were mainly of the European-Mediterranean type (Liorzou et al., 2016). It was at this time that many rose lovers and enthusiasts began to collect and breed roses (Oghina-Pavie, 2015). Many nurseries also specialized in roses. Although previously quite unfamiliar to rose bush cultivation, the city of Lyon quickly specialized in rose breeding (Ferrand, 2015). Within a few decades, the number of new varieties increased tremendously, as well as the phenotypic diversity. It was then common to give Latinized names to the new varieties created though this nomenclature was normally reserved for botanical species (Oghina-Pavie, 2015). This practice brought even more confusion about the already twisted origin of roses (Masure, 2013). Napoleon's wife Josephine de Beauharnais (1763 – 1814) was herself truly fond of roses and gathered thousands of rose bushes corresponding to hundreds of species and varieties at her Chateau de Malmaison. Between 1791 and 1829, the number of rose cultivars available in French catalogues increased from 25 to 2562, possibly thanks to the influence of the collections of Malmaison (Watts, 2009). Josephine sent botanists abroad to bring foreign rose specimens back to France to increase her rose collection. It was also at this time that she commissioned Pierre Joseph de Redouté to paint his well-known watercolors on roses (Redouté and Thory, 1817). Aside from the Chinese tea-type specimens, many other wild rose species were brought back from Asia to Europe such as *R. bracteata* or *R. laevigata* during the 18th century and later on at the turn of the 20th century for *R. davidii*, *R. helenae*, *R. omeiensis* var. *pteracantha*, among other species (Joyaux, 2015). However, hardly any of these newfound species contributed much to the large number of cultivars that were released during the 19th and 20th century but they still represented either curiosities or well suited bushes to natural fences. Wylie (1954) estimated that actually a few ten wild rose species significantly contributed to the vast number of modern roses. During the 18th and 19th century, there were considerable international exchanges of rose material between botanic gardens (Kew Gardens, UK), amateurs' collections (Malmaison and Luxembourg in France) and nurseries (Dupont, Vibert, Cels) with the idea of creating collections that would bring together the diversity of roses in one place.

In 1867, another landmark in rose breeding occurred with the creation of *R.* 'La France'. This cultivar corresponded to a new kind of roses obtain after crossing recurrent blooming specimens with Chinese tea-type roses, therefore pyramiding the most attractive ornamental traits in one cultivar; namely perpetual blooming, fragrance, and abundant flowering. *R.* 'La France' foretold a new era of modern roses, named hybrid tea roses, that are still bred nowadays (Joyaux, 2015). From this year on, the rose industry thrived dramatically and several thousands of rose cultivars were created. The 19th century truly represent the golden age for rose breeding, especially in France (Oghina-Pavie, 2015). Roses were highly prized and were found countless time in literature, from the Romantic poetry (Blake, 1794; Burns, 1834) to children's tales (Grimm and Grimm, 1884; de Saint-Exupéry, 1943), mostly as a symbol of love and perfection. Roses were also used in politics where the red rose became associated with the red flag of socialism (Lee, 1913) (Figure 3I). Nowadays, the range of color, shape, flower type, blooming period, fragrance, is incredible and innovation is still on the move with the introduction of wild type material. For instance, rose breeders Chris Warner and James Sproul succeeded in introducing the mesmerizing macula of *R. persica* into readily crossable material (Heitzler, 2015). This led to the development of new rose varieties with the macula trait such as 'Tigris', 'Euphrate' or the so-called Eye-something lines ('Babylon Eye's', 'Eye of the Tiger') (Figure 5).

Currently, roses represent a major ornamental plant with an important economic value. There are three main outlets for roses: cut flowers, garden rose bushes, and fragrance.

Cut roses - Cut roses represent the biggest market for rose outlets both in terms of value and volume. The total worldwide exportations of cut roses represented $ 3 billion in 2015 (Bouron, 2017). Cut roses are extensively commercialized for life events (weddings, burials) and calendar festivals (Valentine's Day, Mother's Day) (Benoit, 2019). There are several components in the cut roses market: the breeders, the growers, the wholesalers and the retailers. For cut roses, the main breeding companies are historically located in European countries, among others France (Meilland, Delbard) and The Netherlands (Dümmen Orange). However, this European focus is starting to shift as South America (Ecuador, Colombia) and Africa (Kenya) grow in importance and start their own breeding programs (Leus et al., 2018). The current centers of production are in Ecuador (2750 ha), Colombia (2600 ha), for the North American markets, Ethiopia (1200 ha), Kenya (2900 ha) and The Netherlands (238 ha) for European markets, and India (31 000 ha), China (15 000 ha) for Asian-Oceanian markets (Rabobank, 2016). The distribution of production areas in equatorial regions enables to optimize the cultivation of cut roses by (1) providing favorable climatic conditions for an all-year-round rose production and (2) benefiting from cheap labor. Northern markets are therefore supplied at any time of the year. The cut rose industry is highly calibrated and tailored to the production, there is therefore not much room for outstanding ornamental innovations. Important traits are mainly non-ornamentals and concerns easy-to-harvest architecture and shelf life. Indeed, transportation and logistics play an important role in rose industry since

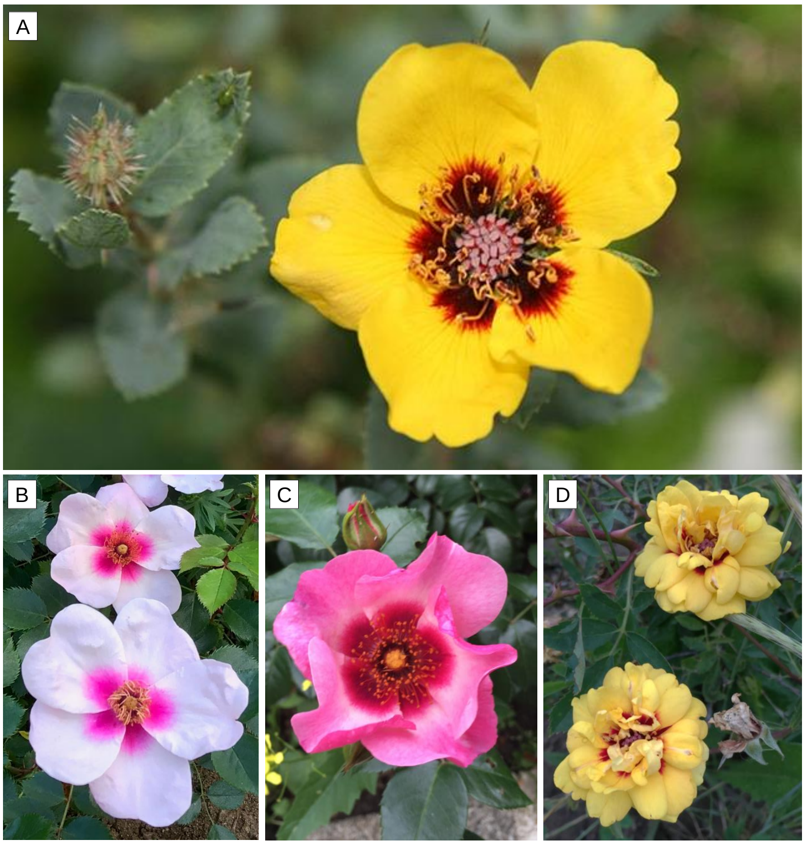


Figure 5: *Rosa persica* and the Hulthemosas. **A**. *Rosa persica* Michx. (Credits: Roses Anciennes en France). **B**. ‘Smiling Eyes’ (Credits: G & JC Spinnler). **C**. ‘Babylon Eyes’$^{©}$ Queen ‘Interybabeucq’ Cov (Credits: K Debray). **D**. ‘Tigris’ (Credits: K Debray)

roses are produced far away from their retail markets. The largest flower auction is located in Aalsmeer, The Netherlands, which country represents a major hub in the rose industry. More than 60% of the worldwide flower market is handled by Dutch auctions. Royal FloraHolland is a Dutch conglomerate of florists and one of the largest auction companies in the world. Between 2009 and 2016, sales were dominated by roses, though these decreased from 2.7 million to 1.8 million pieces. Prices however rose over the same period, from € 0.22/piece to € 0.27/piece. The exportations of cut roses to the EU floral market reached € 1.67 billion in 2016, far ahead from carnation (€ 0.21 billion), chrysanthemums (€ 0.29 billion), lilies (€ 0.12 billion) and orchids (€ 0.08 billion) (Hanks, 2018). It is worth noting the emergence of disposable pot roses as a new market that represent a substantial fraction of the sales with 2.8 million pots sold in France in 2018, mainly during winter, for a total of € 23.2 million (VAL'HOR, 2018).

Garden roses - The garden roses market is far less important than for cut roses and mainly concerns hobby breeding. In 2018, the rose bush market in France represented € 50.2 million (4.7 million rose bushes) while that of cut roses reached € 376.3 million (VAL'HOR, 2018). Over the last ten years, the importance of garden roses has been divided by two in both value and volume (2008: 10.3 million rose bushes; € 95.3 million vs 2018: 4.7 million rose bushes; € 50.2 million) (VAL'HOR, 2018). Breeding for garden roses is mainly tailored to pest resistant plants. Black spot is a major pest causing foliage damages thus reducing the esthetics of the plant (Leus et al., 2018). As more garden roses are grown in containers or tunnels, powdery mildew is increasing in prevalence (Leus et al., 2018). In other regions, like in the USA, other pests such as Rose rosette virus (RRV) are of great importance, along with new emerging pests including rose sawfly, rose midge, rose sludge, thrips (Leus et al., 2018). The ADR (Anerkannte Deutsche Rose) label rewards each year rose cultivars that passed a series of drastic tests during a 3-year trial in eleven German stations (Sieber, 2009). This label is considered to be one of the most difficult to obtain in the world. Since the 1950s and the creation of this label, more than 2000 cultivars have been tested and as of November 2018, 175 cultivars are recognized on the ADR list. In a quest for novelties in rose bushes, rose breeders also focused the innovation on ornamental traits by developing original colors, as those seen in the Hulthemosas (Rowley, 1955) (Figure 5) or the more recent acquisition of a blue rose through genetic engineering (Nanjaraj Urs et al., 2019). Despite the lower commercial importance of garden rose in comparison to cut roses, more research efforts are now carried on cultivation in garden roses, especially in the USA (Leus et al., 2018). Looking at the challenges that garden rose breeding will have to face, it is clear that innovation will play a decisive role in limiting consumer disinterest in garden roses. Given the considerable gain in knowledge acquired over the last decades on rose biology and genetics, the raise of new valuable characteristics, either artificially engineered or captured in the wild pool, could hopefully lend full prestige to garden roses.

Fragrance - Nowadays, the rose attar is essentially produced in Bulgaria and Turkey that supply

80-90% of the world's needs (Kovacheva et al., 2010), and represents a valuable outlet for scented roses (€ 5000-7000/kg). Three to five tons of roses are necessary to produce 1 kg of rose attar or 0.5 kg of rose concrete (Baudino et al., 2013). Given the prohibitive price of rose essential oil, it is almost exclusively used for luxury perfumery. For lower value products, such as flavors in the food industry, similar molecules produced either naturally (through scented-leaved *Pelargonium*) or synthetically are preferred for their lower cost (Verma et al., 2016).

From ancient civilizations to the present day, roses have been an important part of human history. Initially prized for their perfume, roses have established themselves in human culture and symbolism. Through thousand years of selection, people have shaped the five-petal roses into an ever-growing number of more beautiful, fragrant and flowering varieties.

1.3 Description and classification of the genus *Rosa*

1.3.1 What is a rose?

All wild roses belong to the genus *Rosa*, itself a member of the tribe Roseae of the subfamily Rosoideae that is embedded in the large Rosaceae family (Xiang et al., 2016; Stevens, 2017). As a member of the Rosaceae family, the genus *Rosa* shares traits that are specific to this family, such as free petals, numerous stamens, flowers with a 5-fold rotational symmetry, alternate leaves, presence of stipules (except in *R. persica*) (Malécot, 2015). In addition, the genus *Rosa* has its own characteristics that distinguish it from other genera of the family Rosaceae. In particular, *Rosa* species correspond to sarmentose shrubs, having an urn-shaped floral receptacle, numerous free carpels, and odd-pinnate leaves (except *R. persica*) (Malécot, 2015) (Figure 6). There is a great diversity in color, shape, size, and aspect of organs across the different species. Here, we review the main characteristics of each organ as well as the diversity found between *Rosa* species. General descriptions are mostly based on the review done by Masure (2013), personal observations, and extra references are mentioned to complete the descriptions.

Architecture, stems and roots

Species of the genus *Rosa* correspond to woody perennial shrubs that can be erect, climbing or trailing. Dimensions of wild rose bushes range from the small briar of *Rosa persica*, barely higher than 60 cm, to the impressive vines of *Rosa gigantea* that can exceed 15 meters and occupy the entire cyme of a large tree. The stems can be vigorous, erect, or conversely hail and soft, trailing or climbing if they find a support nearby. Stems are usually green during the first years of growth and turn grey/brown when older (Figure 7). Many young stems secrete bloom, a thin waxy film that disappear on older stems. It is worth noting the particularity of species from the *Platyrhodon* subgenus which have stems with a bark that peel off (Figure 7G). These species look like trees of a few meters high when mature. Although some species are glabrous, stems usually include sharp

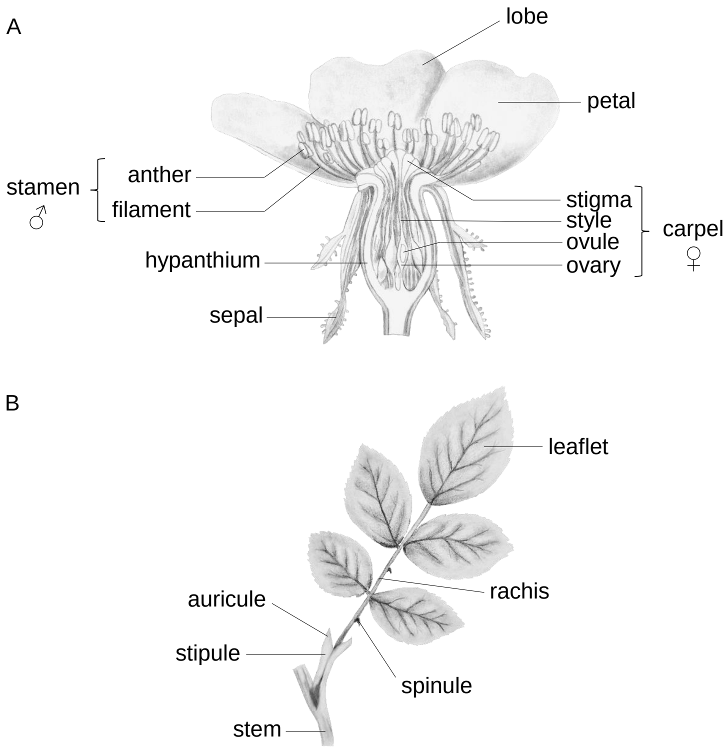


Figure 6: Morphology and anatomy of wild rose. **A**. Cross-section of a flower. **B**. Typical odd-pinnate leaf. Lead pencil reproductions by K Debray.

structures, incorrectly called thorns while they rather correspond to prickles since they derive from the stem sub-epidermis and are not vascular, unlike true thorns which are modified stems. Rose prickles generally look like sickle-shaped hooks but there exists a large diversity in prickle shapes, and they can be straight, curved or hooked, thin or wide, cylindrical or flattened and allow reliable identification of certain species (Figure 7). Prickle colors range from white to ruby red, with all the shades of green, grey and purple. Rose prickles have a protective role against grazing animals for all low shrubs growing in open areas (dune, meadow) and may indirectly help rose bushes grabbing the surrounding plants in order to grow upright and better access light in more wooded places, especially in forest edges. Sometimes, prickles get mixed with bristles resulting in a dense prickling that may also help coastal species limit erosion of their substrate by retaining windblown sand thus preventing root exposure. Wild roses have a tap root system and some species sucker in abundance from roots and underground stems, forming dense colonies (*R. rugosa*, *R. spinosissima*).

Leaves

Leaves are inserted alternately in a spiral along the stem and are generally 5 to 15 centimeters long, although their size varies considerably between species, ranging from 1.5-3 cm for *R. persica* to 20 cm for some varieties of *R. longicuspis*. The leaves are always odd pinnate and display stipules (Figure 6B), except for leaves of *R. persica* that are simple, sessile and lack stipules (Figure 8B).

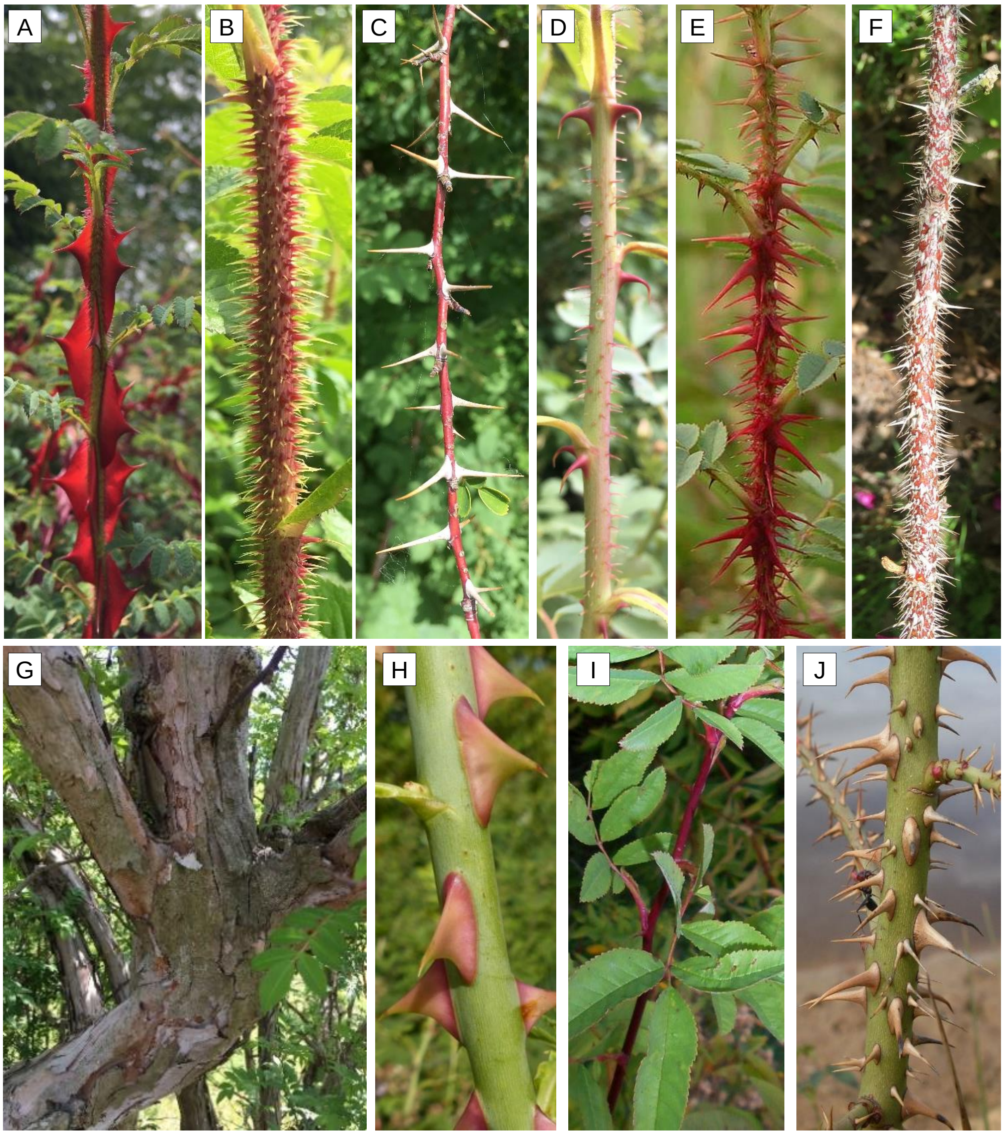


Figure 7: Diversity in stems. **A**. Stem of *R. omeiensis* var *pteracantha* with large and deep red prickles. **B**. Densely prickled stem of *R. maximowicziana*. **C**. Long prickles and thin stem in *R. forrestiana*. **D**. Hooked prickles with bristles in *R. fedtschenkoana*. **E**. Young stem with red prickles in *R. spinosissima*. **F**. Old stem of *R. rugosa*. **G**. Large trunk-like stem of *R. roxburghii* var *hirtula*. **H**. Prickles on young stem of *R. canina*. **I**. Nearly glabrous stem of *R. palustris*. **J**. Stem of *R. rubiginosa*. Credits: K Debray, except E (Velela), H (Easy Wild Flowers), I (Salicyna), J (John Tann).

Stipules are mostly adnate, meaning that they are fused to a large part of the petiole. Free parts of stipule are called auricles and can be more or less divergent from the petiole axis. A serrated pattern can often be observed on the stipule margins and stipules can be glandular. As for stipule, leaflet margins can be serrated, and small prickles (spinules) can grow underneath the rachis (Figure 6B). On average, rose leaves are composed of 5 to 9 leaflets, although it varies between species. For instance, *Rosa glutinosa* and *R. banksiae* usually have 3-5 leaflets per leaf while *R. roxburghii* and *R. omeiensis* can display up to 17 leaflets (Cuizhi and Robertson, 2003). Most wild roses are deciduous although some species are evergreen or nearly so, such as *Rosa sempervirens* (Klastersky, 1968). The foliage also offers a wide range of colors and shapes (Figure 8A, C and D). Some foliage are light green, others grey, deep red or even bluish-grey. Some leaves are smooth and shiny while others are rough or felted, often glandular and even sometimes aromatic. Leaves morphology generally reveals the environmental conditions in which roses adapted. In dark and humid habitats, plants develop large, often glabrous leaves and stems with very few prickles (ex. *R. laevigata*). On the contrary, roses growing in dry conditions limit evapotranspiration being equipped with reduced leaves, sometimes hairy, with stems armed with a dense and developed protective thorny system (*R. persica*, *R. minutifolia* (Figure 8E), (Baldwin et al., 2012)).

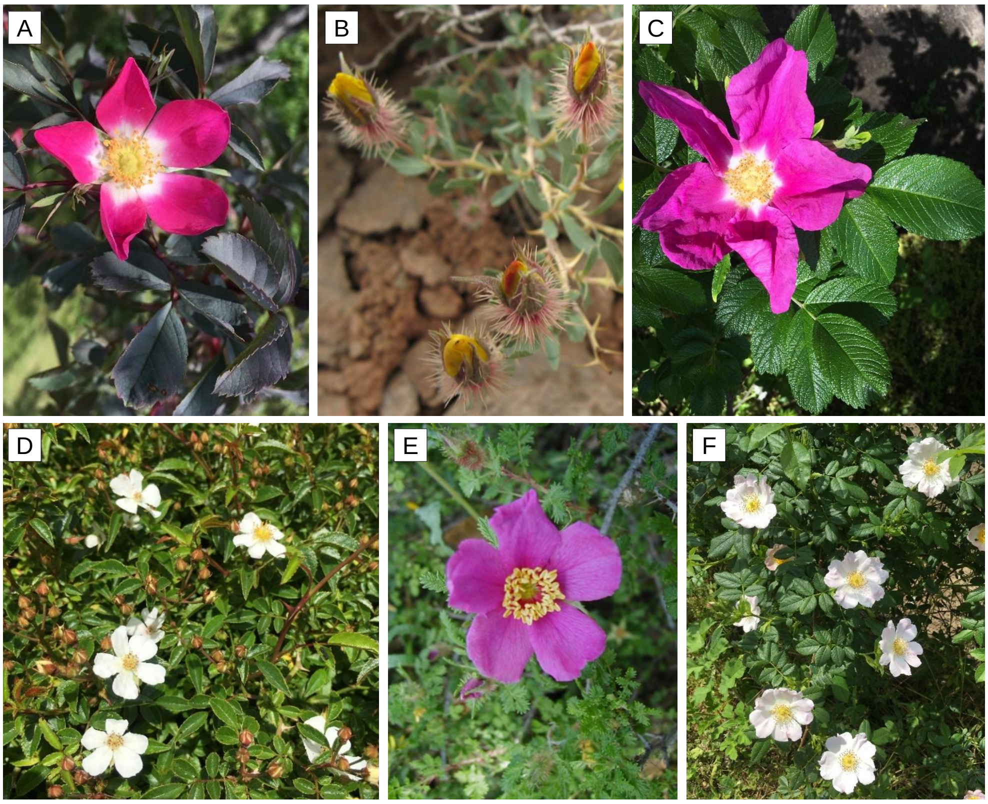


Figure 8: Diversity in leaves and leaflets. **A**. Dark leaves of *R.* textitglutinosa. **B**. Simple leaves of *R.* textitpersica. **C**. Rough leaves of *R. rugosa*. **D**. Shiny leaves of *R. wichurana*. **E**. Tiny leaves of *R. minutifolia*. **F**. Leaves of *R. canina*. Credits: K Debray, except B (Arghiyan) and E (Charles E. Jones).

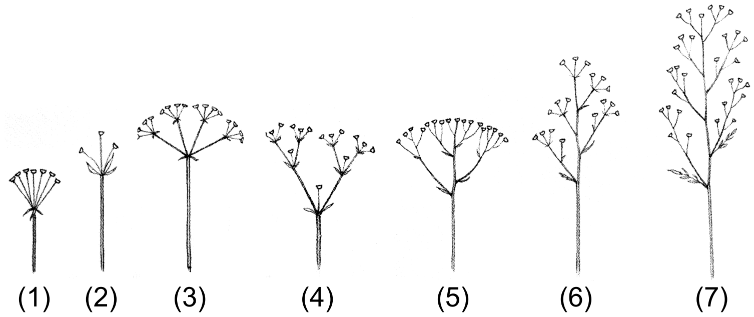


Figure 9: Diversity in inflorescences. (1) Umbelliform cyme, (2) Cyme, (3) Compound umbel, (4) Compound cyme, (5) Cymose corymb, (6) Cymose panicle, (7) Compound panicle. Reproduced from Masure (2013).

Flowers

Inflorescences generally gather 2-3 flowers in a bouquet called cyme. Solitary flowers are quite uncommon but still occur. Several species from section *Synstylae* display large multi-flowered inflorescences: umbelliform cyme, cymose corymb, panicle (Figure 9). The denser bouquet, the smaller the flowers are. Wild rose flowers are actinomorphic, rotaceous (bowl-shaped) flower and carry five petals except *R. sericea* that usually has only four (Cuizhi and Robertson, 2003) (Figure 10B). Regarding flower diameter, it varies from 2 cm for *R. cymosa* (Figure 10F) to 12 cm for *R. gigantea* (Figure 10I). Two distinct lobes are present on the apex of each petal (Figure 5A) and the range of colors varies from white to pink, sometimes crimson or yellow, with all shades in between (Figure 9). The basal part of each petal is generally slightly lighter than the rest of the blade, except for *R. persica* that displays a macula, ie dark petal base versus light blade. Pollinating insects seem to be more attracted by contrast than by the colors themselves (Hirota et al., 2018), especially since most insects can perceive ultraviolet light (Primack, 1982). Such a color contrast on the petals of *Rosa persica* is likely the result of coevolution with local pollinating insects (Heitzler, 2015) (Figure 10H). Underneath the corolla, the calyx is made of five sepals (four in the case of *R. sericea*) that can be tapered, lanceolate or foliate. In some species, sepals can be highly glandular (*R.* × *centifolia* var. *muscosa*) giving the moss character. The time that sepals remain on flower and then on fruit varies from one species to another, as well as their orientation during fruit ripening. Persistent or deciduous, deployed, erect or reflected, are all criteria that help identify species. Flowers are mostly hermaphrodite (but *R. setigera* Michx. is criptically dioceious (Kevan et al., 1990)) and contain both staminate and carpellate parts (Figure 6A). There are numerous stamens, and anthers are generally bright yellow in young flowers and turn brown after few days of opening. Anthers are composed of two pollen sacs which contain pollen. The numerous carpels are free (i.e. not fused in a single ovary) but have free (most of the genus) or fused styles (Section *Synstylae*) and are enclosed within a deep urceolate hypanthium that turns fleshy when mature. The carpels develop into achenes after pollination (MacPhail and Kevan, 2009).

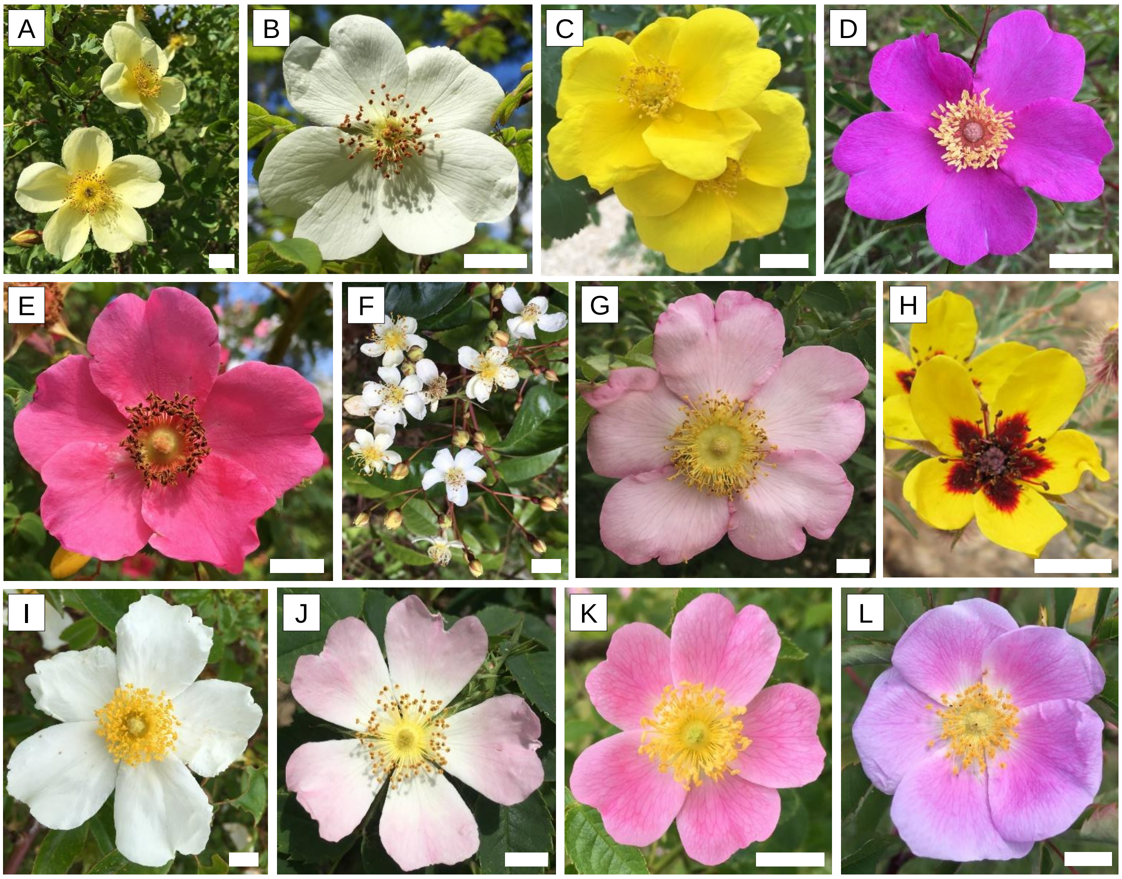


Figure 10: Diversity in flowers. **A**. Yellow flowers of *Rosa hemisphaerica* var *rapinii*. **B**. 4-petal flower of *R. sericea*. **C**. Brigh yellow flower of *R. foetida*. **D**. Brigh purple flower of *R. acicularis*. **E**. Fuschia flowers of *R. moyesii*. **F**. Tiny flowers of *R. cymosa*. **G**. Large flower of *R. roxburghii* var *hirtula*. **H**. Eyespotted flower in *R. persica*. **I**. Giant flower in *R. gigantea*. **J**. Flower of *R. canina*. **K**. Small flower of *R. bella*. **L**. Large petal flower of *R. davurica*. Credits: K Debray, except H (Arghiyan). White bar indicates a length of 1 cm.

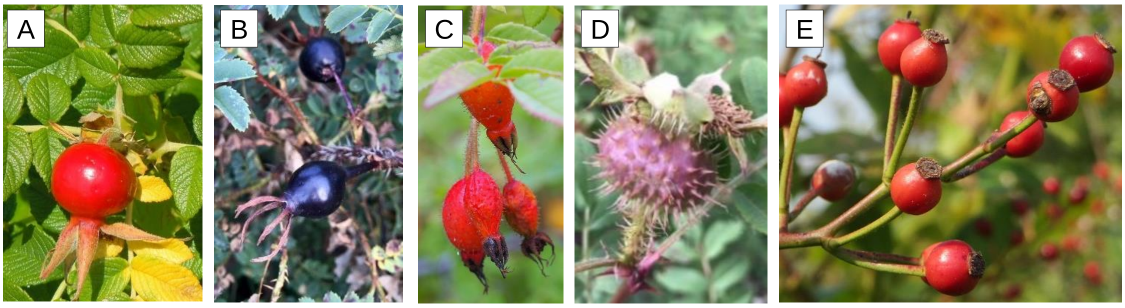


Figure 11: Diversity in rosehips. **A**. A round and fleshy hip of *Rosa rugosa*. **B**. Black hips of *R. spinosissima*. **C**. Dangling hips of *R. pendulina*. **D**. A chestnut-like hip of *R. roxburghii*. **E**. Grape of small rounded hips of *R. helenae*. Credits: K Debray, except A (VFClark), C (Agnieszka Kwiecień, Nova), E (AnRo0002).

Fruits

Wild rose accessory fruits are called hips (FR: cynorhodon). Most of the time, rose hips are red when mature although yellow to black hips can be observed in some species (Figure 11). Rose hips are generally globular or subglobular, but there are also piriform, ureolate, top-, turbine-, bottle- or amphora-shaped fruits. Smooth and glossy, hispid or glandular, rose hips offer a nice overview of the diversity that can be observed within the genus *Rosa* (Figure 11). The fleshy layer that corresponds to the hypanthium carries on its internal surface tens of achenes, each one actually represents a true fruit. Each achene is the result of one carpel transformation. The ovary wall alters in pericarp and a seed coat separates the seed from the pericarp. Thin but stiff hairs derived from the hypanthium and achene pericarps envelop the achenes.

Chromosome and genomic considerations

The basic chromosome number, ie the number of chromosomes in a single non-homologous set of chromosome, is 7 (x = 7) (Täckholm, 1920; Hurst, 1925). About half of wild rose species are diploid (2n = 2x = 14) while the remaining species are polyploid. Euploidy predominates with almost all levels of even and odd ploidy levels ranging from triploid (2n = 3x = 21) to decaploid (2n = 10x = 70). Aneuploidy has never been reported yet in wild rose species but has already been recorded in experimental progenies (Rowley, 1960). Monoploid genome size (1Cx) of roses varies depending on the species but is comprised between 0.37 pg for *R. zhondianensis* and 0.89 pg for *R. brunonii* (Yokoya et al., 2000; Roberts et al., 2009; Jian et al., 2014). Recent genome sequencing initiatives of *R.* ‘Old Blush’, a putative diploid hybrid within the section Chinenses (Meng et al., 2011), revealed a haploid genome size of approximately 500 Mb encompassing nearly 40,000 protein coding genes (Raymond et al., 2018; Hibrand Saint-Oyant et al., 2018). The authors found that about 63-68% of the genome sequence is composed of transposable elements. A strong synteny between the woodland strawberry (*Fragaria vesca*) and *Rosa* has also been highlighted (Hibrand Saint-Oyant et al., 2018). In addition, no major recent whole genome duplications has been observed in the reference genome (Hibrand Saint-Oyant et al., 2018).

1.3.2 Mechanisms of reproduction and diversification

Sexual reproduction

Wild rose flowers usually bloom in the morning during spring, from late April to early July, in their natural habitat of the northern hemisphere. The bowl shape of wild rose flowers allows many different insects (Apidae and Syrphidae) to access the flower thus entomophily is the preferred way of pollination in wild roses (MacPhail and Kevan, 2009). Although the flowers are devoid of nectar, attractive petal colors, scent and abundant production of pollen are all rewards for potential pollinators. Wild rose flowers are mostly hermaphrodite and sexual reproduction involves a seed plant (♀) and a pollen plant (♂). Autogamy is not tolerated much since self-pollinating causes inbreeding depression and a lower fertility, although this varies depending on the species and its ploidy levels (MacPhail and Kevan, 2009). Indeed, diploid species seem to be less prone to autogamy than polyploid species (MacPhail and Kevan, 2009). Allogamy is mainly favored through self-incompatibility mechanisms that prevent fertilization from occurring if pollen and stigma belong to the same flower or flowers of the same plant (Debener et al., 2010; Caser, 2017). In addition, cryptic dioecy has also been reported in *Rosa setigera* (Kevan et al., 1990), meaning that male- and female-type plants are separated thanks to either anther or stigma sterility which therefore facilitate crosses between different genotypes.

In the simple case of diploid progenitors, each parent usually produces haploid gametes which merge with gametes from the opposite sex and form a diploid embryo after pollination. However, there are some variants to this classic scheme that make it possible to extend the mix of genetic diversity and speed up the pace of standard evolution. Indeed, wild roses succeeded in spreading across the northern hemisphere, conquering very diverse habitats, suggesting highly plastic genomes that favor genetic exchanges and therefore genetic diversity. The range of sexual reproductive mechanisms that wild roses have developed through years of evolution may partly explain their evolutionary success (Ritz and Wissemann, 2011). First, interspecific barriers are loose within the genus and fertile hybridizations between different species are quite common (Kellner et al., 2012b; Andersen et al., 2016; Vaezi et al., 2019). Interspecific crosses allow favorable alleles to be combined more quickly, each new allele being put to the test of purifying selection (Alix et al., 2017). Heterotic offspring may better resist rapid environmental changes. In addition to the ease of hybridize, wild roses can combine different ploidy levels. Therefore, genetic exchanges are much more significant when hybridization includes or results in polyploid species because more alleles are involved (Cole and Melton, 1986). Although interspecific crosses might reduce fertility in plants, whole genome duplication after pollination seems to be an efficient solution to restore fertilization and stabilize hybrids between distantly related species (de Wet, 1971; Fougère-Danezan et al., 2015). Another way of polyploid formation in plants corresponds to crosses involving unreduced gametes which seems more common in roses (Ritz and Wissemann, 2011; Rani et al., 2013; Herklotz and Ritz, 2017), and can be promoted by environmental conditions such as temperature (Pécrix et al., 2011). For instance, triploid plants may originate from a diploid-diploid cross involving one unreduced gamete (2n) and one normally reduced gamete (1n). By extension, tetraploid plants may have diploid progenitors as long as both parental lineages have produced unreduced gametes (2n), but could also arise from a triploid plant with unreduced gametes (3n) and a normally diploid plant with haploid gamete (1n) (Zlesak, 2009). Virtually all ploidy levels exist in wild roses from 2x to 10x and even multiploid species (Lewis, 1959), ie species that include specimens with different ploidy levels, were reported. There are scarce references on polyploid origins of most rose species and sections. From cytogenetic analyses, the frequent observation of bivalent chromosomes (disomic inheritance) in polyploid specimens (Hurst, 1928) might indicate the predominance of allopolyploidy, ie polyploidization involving different species (Figure 12), over autopolyploidy, ie polyploidization involving same-species specimens (Figure 12). However, the lower amount of multivalent pairings, generally associated with polysomic inheritance and autopolyploidy (Ramsey and Schemske, 2002), does not mean that they have never occurred in *Rosa*. Indeed, more ancient autopolyploids tend to behave like diploids and show bivalent instead of multivalent pairing (Le Comber et al., 2010). Autogamy is also more common as ploidy levels increase (MacPhail and Kevan, 2009) since more allele are present and may be sufficient to circumvent self-incompatibility mechanisms (Bourke et al., 2017).

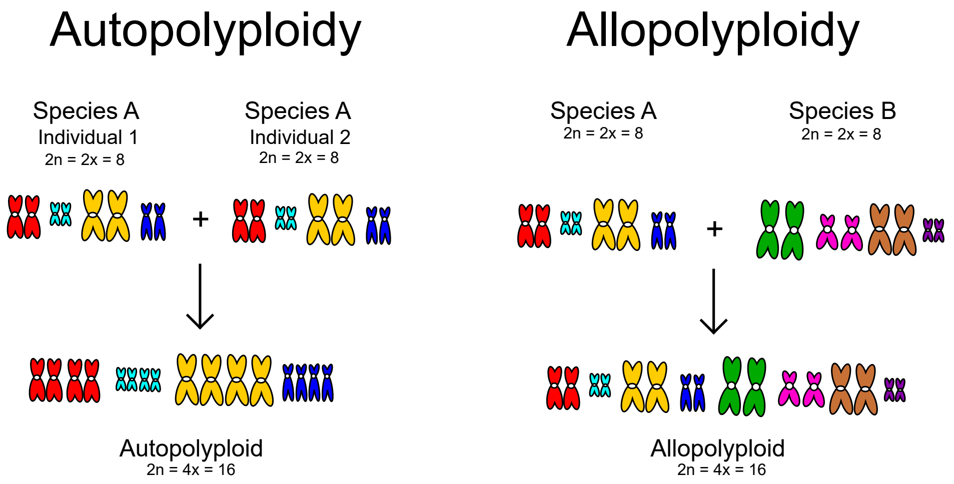


Figure 12: Autopolyploid versus allopolyploid formation. Autopolyploid taxa arose from an intraspecific cross while allopolyploid taxa are the result of an interspecific hybridization with whole genome doubling. In recent autopolyploid, quadrivalent pairing is often observed while bivalent pairing is usually maintained in allopolyploid. Note that the classification into these two categories is not always straightforward (see text).

An outstanding example of mixed, recurrent hybridization and polyploidization lies in the polyploid section *Caninae* (4x, 5x, 6x). Unbalanced meiosis has been observed during megasporogenesis in both carpels and anthers and exemplifies the so-called "*Canina* meiosis" (Täckholm, 1920; Nybom et al., 2004). *Caninae* genomes are composed of two sets of seven bivalent chromosomes that segregate during meiosis and 14, 21 or 28 univalent chromosomes that do not segregate (Täckholm, 1920) and are typically inherited from the egg cell (Werlemark and Nybom, 2017) (Figure 13). Therefore, the androecium produces haploid gamete (1n = 1x = 7) whereas the gynaeceum provides tri- (1n = 3x = 21), tetra- (1n = 4x = 28), or pentaploid (1n = 5x = 35) gametes depending on the initial ploidy level of the maternal parent (Werlemark and Nybom, 2017). Sometimes, the female gamete is totally unreduced leading to an ovule of the same ploidy level as the female plant that is able to cross with haploid pollen therefore resulting in an embryo with an increased ploidy level. The unbalanced *Caninae* meiosis results in a matroclinal inheritance of traits, limiting the distinction between hybrids and their maternal line (Werlemark and Nybom, 2017).

Hybridization events associated or not with polyploid shifts are very common in *Rosa* and virtually all combinations are possible. This conveys the idea that both hybridization and polyploidy are major driving forces in evolutionary history of *Rosa*.

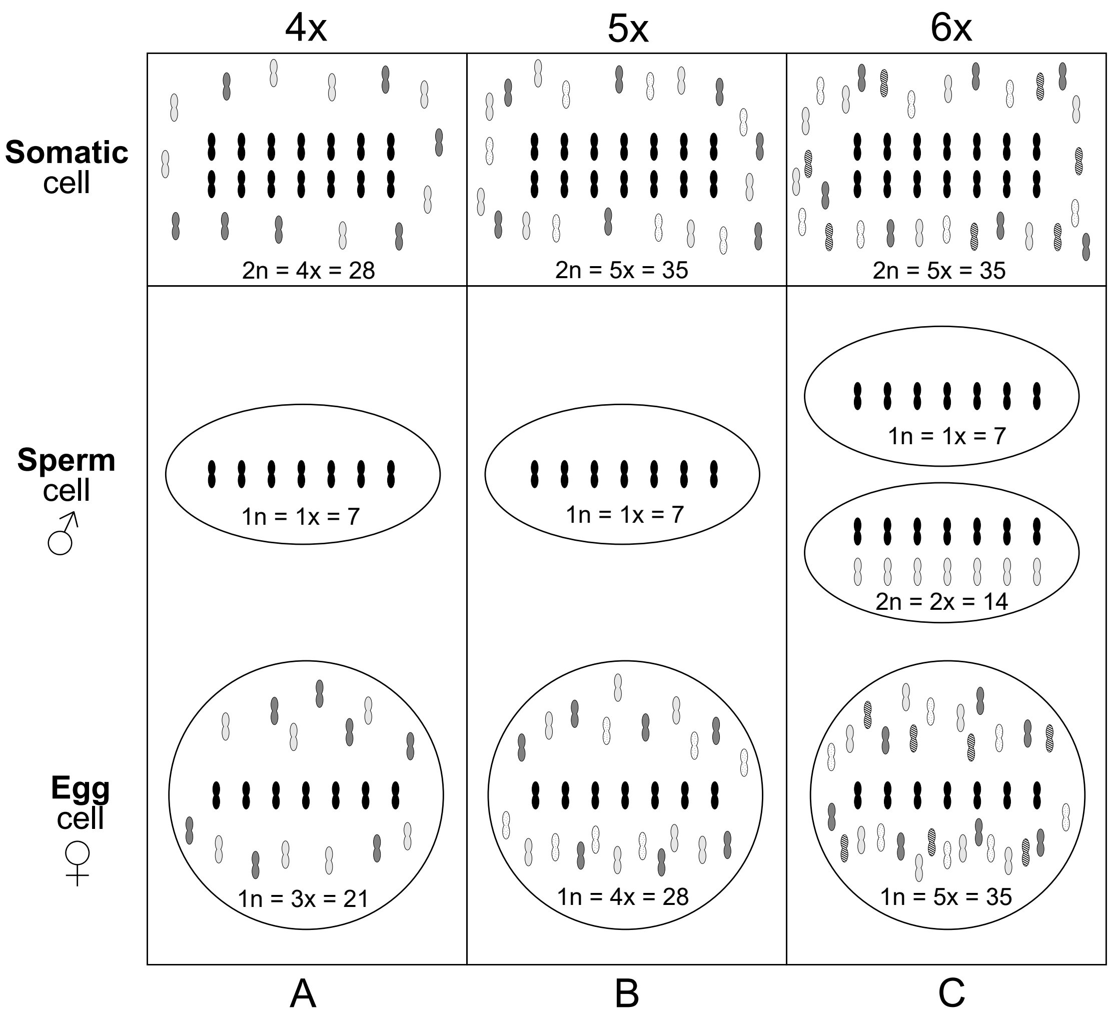


Figure 13: A schematic representation of gamete formation in dogroses. Each somatic cell has 2 septets of bivalent chromosomes (black) and 2, 3, 4 septets of univalent chromosomes in 4x, 5x, 6x individuals respectively (light gray, dark grey, dotted, striped). The *Caninae* meiosis corresponds to an unbalanced meiosis where the sperm cell gets one septet of bivalent chromosomes (1n = 1x = 7) (sometimes 2 (2n = 2x = 14) in hexaploid *R. micrantha*) and the egg cell gets the other septet of bivalent chromosomes plus the septets of univalent chromosomes. **A**. Meiosis in tetraploid dogrose. **B**. Meiosis in pentaploid (most common) dogrose. **C**. Meiosis in hexaploid dogrose. Adapted from (Ritz and Wissemann, 2011).

Asexual reproduction

Aside from sexual reproduction, wild roses can propagate through asexual reproduction with two main strategies. The first one corresponds to vegetative reproduction. Several rose species spread locally using underground basal shoots such as *R. gallica* (Joyaux, 2015), *R. rugosa* (Cuizhi and Robertson, 2003), *R. spinosissima* (Boyd, 2015) or *R. persica* (Heitzler, 2015). For the latter one, developing root sprouts from a perennial deep-rooted system may be an adaptive trait to arid environment where heat can be detrimental to young and fresh shoots. The second asexual reproductive system in wild roses corresponds to apomixis (agamospermy), ie clonal reproduction through seeds. It has so far only been observed in the *Caninae* complex and in few species such as *R. rugosa* (Dobson et al., 1999) and *R. virginiana* (MacPhail, 2007), although agamospermy seems to be facultative and represents the least effective means of producing hips among all the breeding systems that were examined (MacPhail and Kevan, 2009).

1.3.3 The thorny notion of species and species delineation in *Rosa*

Delimiting species boundaries is fundamental to elucidate the organization of biodiversity and a reliable definition of species truly matters in many fields of biology, from conservation to ecology. Currently, species are given a binomial name with the first part being the generic name and a second part corresponding to the specific epithet. This binomial nomenclature has been first established by Carl Linnaeus in *Species Plantarum* (1753), and has been carried on for its practicality. The Linnaean classification more generally corresponds to a nested system where each organism belongs to a determinate entity. In his vision, species, which were created by God, are immutable and the origin of species is purely theological (McGregor Reid, 2009). As soon as the Linnaean system was adopted, it was criticized for its inflexibility, especially by the Comte de Buffon (Sloan, 1976; Hoquet, 2007) who latter inspired a new generation of naturalists, namely de Lamarck (Björklund, 2019) and Darwin (Darwin, 1866). During the 19th century, the emergence of theories of evolution revolutionized the species concept (Bowler, 2003; Larson, 2006). These new theories alleged that species are born, transform, then die or give birth to new species. The species are thus related to each other through a global evolutionary pattern. From this period onwards, species are understood in their spatial and temporal dimension and evolution corresponds to a genuine process that makes the species concept illusory because species are unstable. Indeed, the boundary between species and variety blurs since varieties can be defined as emerging species (Darwin, 1866). These 19th century theories recognized the importance of intraspecific variability and the challenge was to reconcile the species concept with evolutionary theories. During the 20th century, the definition of species recognizes its temporal and changing dimension. Species is therefore defined as a set of individuals that intercross to give a fertile progeny from one break to the other along the genealogical flow (Mayr, 1942; Huxley, 1943; Stebbins, 1950).

Numerous species concepts have been recognized, with some very sophisticated (Häuser, 1987; De Queiroz, 2007; Hausdorf, 2011). While some authors claim for a Phylogenetic Species Concept that emphasizes on monophyletic lineages as an undeniable evidence of species delineation (Hennig, 1966; Donoghue, 1985; Baum, 1992), others argue that the notion of reproductive isolation is a key and ultimate standard for recognizing species (Biological Species Concept) (Mayr, 1942). While many theories exist on the species concept, in practice the distinction of one species in relation to another is essentially based on three criteria: morphology, phylogeny and biology (Wiens and Penkrot, 2002; Spooner, 2016).

Wild roses encompass a large diversity of morphology and habitats which could be a priori a nice asset to distinguish between species as long as each species can be defined by its morphology and habitats. However, while the number of morphological traits for species identification seems to be large, it becomes scarce when dealing with close species relationships (Kellner et al., 2014). Moreover, some species encompass a large diversity of habitats so that their ubiquity cannot help discriminate species. Even worse, some hybrid species may have occurred spontaneously in different places from common parental species, though their lineages may be different. These are called polytopic species and examples were found in section *Caninae* (Herklotz and Ritz, 2017) and can be supposed in *R. spinosissima* given that it occurred in very different habitats and encompasses a large intraspecific diversity. The range of reproductive mechanisms in the genus *Rosa* clearly brings some mess to the matter. Hybridization and polyploidy greatly complicate clear identification of species, because heterosis with or without whole genome duplication scramble morphological traits, with sometimes unequal contribution of the parental lines in the case of the "*Caninae* meiosis". The traditional modern species concept states that species corresponds to the largest group of organisms with similar traits in which individuals can produce a fertile offspring that resemble its parents from one speciation event to the other along the genealogical flow. This suggests that both species relationships and evolution can be represented using a bifurcating structure, namely a phylogenetic tree. Given the ability of roses to spontaneously hybridize to give birth to new species, it seems clear that this traditional species concept is quite inappropriate. In this way, new concepts arose to better describe the complexity of species evolution. For instance, the klepton concept has been developed as an intermediate term between the species and generic name and serves to indicate the result of hybridogenesis (Dubois, 2011). It has been used by some authors in *Rosa* to describe groups in section *Caninae* (Mercier, 2014), where it is difficult to clearly define species. The term is borrowed from zoology and applies to polyploid species that "steal" genomes of other taxa during reproduction without mixing them with their own genome (Dubois, 2011). Each taxon consists of lines reproducing according to the *Caninae* meiosis with a stable genetic part (univalent chromosomes) and a mobile genetic part (bivalent chromosomes). The former is maternally inherited and the latter is used during sexual reproduction and is interchangeable with many other taxa of the genus (Dubois, 2011). Other *Rosa* species can be seen as mayrons, that is to say species s.s. including interconnected lines able to reproduce sexually thus showing high heterogeneity level due to permanent genetic mixing (Mercier, 2014).

A common issue when dealing with *Rosa* species is the correct identification of samples. In addition to their ability to cross between species even at different ploidy level, correct identification of *Rosa* hybrids is sometimes further complicate due to unbalanced trait inheritance (Werlemark and Nybom, 2017). In addition to the natural and intrinsic complexity of the genus *Rosa*, there is the artificial complexity added by the hand of man. Indeed, after centuries of selection and description, there is a lot of confusion within the genus which probably partly explains the incredible number of synonyms that some species can have (Brumme et al., 2013). Therefore, it seems clear that misidentification may occur frequently either during wild sampling if different *Rosa* species populations overlap or in botanical garden if the specimen were grown from seeds.

Many attempts to find a consensus on *Rosa* classification were issued during the last centuries. While each of them bear some flaws, they still represent a valuable ground on which to further improve the genus *Rosa* classification, in this desire to combine both species definition and evolution.

1.3.4 Classification using morphological traits and descriptions

Before the advent of molecular phylogenetics, *Rosa* classification relied solely on morphological characteristics. Here, we review a brief history of several attempts to classify wild roses based on morphological traits. For a recent and detailed review on *Rosa* taxonomy, please refer to (Tomljenovic and Pejic, 2018).

The 19th century marks the transition between artificial classification and scientific (natural) classification (Mayr and Bock, 2002). In the artificial classification, species were described and classified according to more or less subjective criteria, often linked to putative medicinal aspects. Notably, two Greek botanists, namely Theophrastus (B.C. 371 − 287) and Dioscorides (40 − 90 A.D.) set the ground for plant classification. Their works were later resumed during the 16th century stimulated by the Age of Discovery, especially by extending the knowledge about medicinal uses of each plant described. This led to the publication of many state-of-the-art books such as The Herball or Generall Historie of Plantes (1597) by John Gerard (c 1545-1616), largely based on works done by the Flemish botanist Dodoens (1517 – 1585). John Gerard devoted 13 pages to roses making a distinction between musk roses and wild roses. Description of roses are generalized and include different morphological traits such as color and scent of flowers, shrub heights and petal/sepal number. This is followed by numerous less artificial classification of plants, based on morphological traits from fruits and seeds (Andrea Cesalpino (1519-1603)) or flower corolla (August Rivin (1652-1723)) (Tomljenovic and Pejic, 2018). In his famous *Species plantarum* (1753), Linnaeus described 12 species of roses using the binomial nomenclature (*R. cinnamomea*, *R. eglanteria*, *R. villosa*, *R. canina*, *R. spinosissima*, *R. centifolia*, *R. alba*, *R. gallica*, *R. indica*, *R. sempervirens*, *R. pendulina*, *R. carolina*) and added *R. pimpinellifolia* in his edition of 1759, but did not mention any classification. In the later version, Linnaeus divided the genus *Rosa* in two groups: *germinibus subglobosus* and *germinibus ovalis*. Later on in 1815, Desvaux published a paper on roses in France, in which he suggested a division in two groups based on the presence of free or joint stylus (Tomljenovic and Pejic, 2018).

Like cultivated roses, wild roses became very popular from early 19th century onwards. Many botanists and naturalists showed interest in this genus and proposed scientific classifications based mainly on the sharing of common morphological characteristics. In this way, de Candolle divided the genus Rosa into 11 sections (*Synstyleae*, *Rubigineae*, *Gallicanae*, *Chinenses*, *Cinnamomeae*, *Hebecladae*, *Pimpinellifoliae*, *Villosae*, *Centifoliae*, *Caninae* and *Eglanteriae*) as related in Seringe (1823). Lindley (1820) proposed some arrangements and also suggested 11 sections (*Simplicifolia*, *Feroces*, *Bracteatae*, *Cinnamomeae*, *Pimpinellifoliae*, *Centifoliae*, *Villosae*, *Rubiginosae*, *Caninae*, *Synstylae* and *Banksianae*) and detailed the description of 76 species. It is worth mentioning that many *Rosa* classification attempts were made by West European botanists, some of whom were quite concerned about classifying dog roses. Either ordered under the section *Cynorhodon* (Burnat and Gremli, 1886) or *Caninae* (Christ, 1873), these botanists further distinguished subsections *Vestitae*, *Rubigineae*, *Tomentellae*, *Trachyphillae* (not in Burnat and Gremli (1886)) and *Caninae*. Crépin (1889, 1891) divided the genus *Rosa* into 15 sections (*Synstylae*, *Stylosae*, *Indicae*, *Banksiae*, *Gallicae*, *Caninae*, *Carolinae*, *Cinnamomae*, *Pimpinellifoliae*, *Luteae*, *Sericeae*, *Minutifoliae*, *Bracteatae*, *Laevigatae* and *Microphyllae*). In the same view as Crépin (1889, 1891) and Boulenger (1924, 1933, 1935, 1936) to simplify and reduce the number of *Rosa* species, Rehder's classification (1940) reported one hundred to 200 species in temperate and subtropical regions of the Northern Hemisphere. He divided the genus into four subgenera: *Eurosa* (69 species), *Hulthemia* (1 species), *Platyrhodon* (1 species), and *Hesperhodos* (1 species). His subgenus *Eurosa* contains 10 sections (*Pimpinellifoliae*, *Gallicanae*, *Caninae*, *Carolinae*, *Cinnamomae*, *Synstylae*, *Indicae*, *Banksianae*, *Laevigatae* and *Bracteatae*) (Figure 14 and 15). Later on, arrangements were proposed, notably the exclusion of *R. persica* (subg. *Hulthemia*) was commented by de la Roche et al. (1976) although it was already an issue in the 19th century (Dumortier, 1824). de la Roche et al. (1976) further suggested to rename section *Cinnamomeae* in *Rosa* and *Indicae* in *Chinenses*. Wissemann (2003a) further subdivided the section *Caninae* in six subsections (*Trachyphyllae*, *Rubrifoliae*, *Vestitae*, *Rubiginae*, *Tomentellae*, *Caninae*) based on the works of Christ and Crépin. The main traits used to distinguish between species were among others branches, shoots, prickles, setae, leaf glands, branch hairiness, pedicels and orifices, stipules, leaves, leaf color, leaflets shape, flower and blossom, bracts, sepals, and ovary (Rehder, 1940). Hereafter, we detailed the general features that help distinguish between *Rosa* subgenera and sections based on the descriptions from Masure (2013).

Sub-genus *Hulthemia* (Dumort.) Focke

It contains only one species (*Rosa persica* Michx.) originating from Central Asia (Iran, Afghanistan, Iraq, Kazakhstan, near the Caspian Sea and the Aral Sea, on stony and salty soils, Siberia, China). Because of its singularity, the inclusion of *R. persica* to the genus *Rosa* was questioned by some authors. It is a dwarf shrub, bushy, raking, suckering, not exceeding 30 to 50 (90) cm. Young yellow and smooth shoots, turning brownish yellow, glabrous or pubescent, particularly thin, erect or arched, with small yellow, translucent spurs on the young shoots, tapered, curved or hooked, implanted in pairs under the leaves. The leaves are sessile, 1 to 3 cm long, simple, glaucous, ellip-

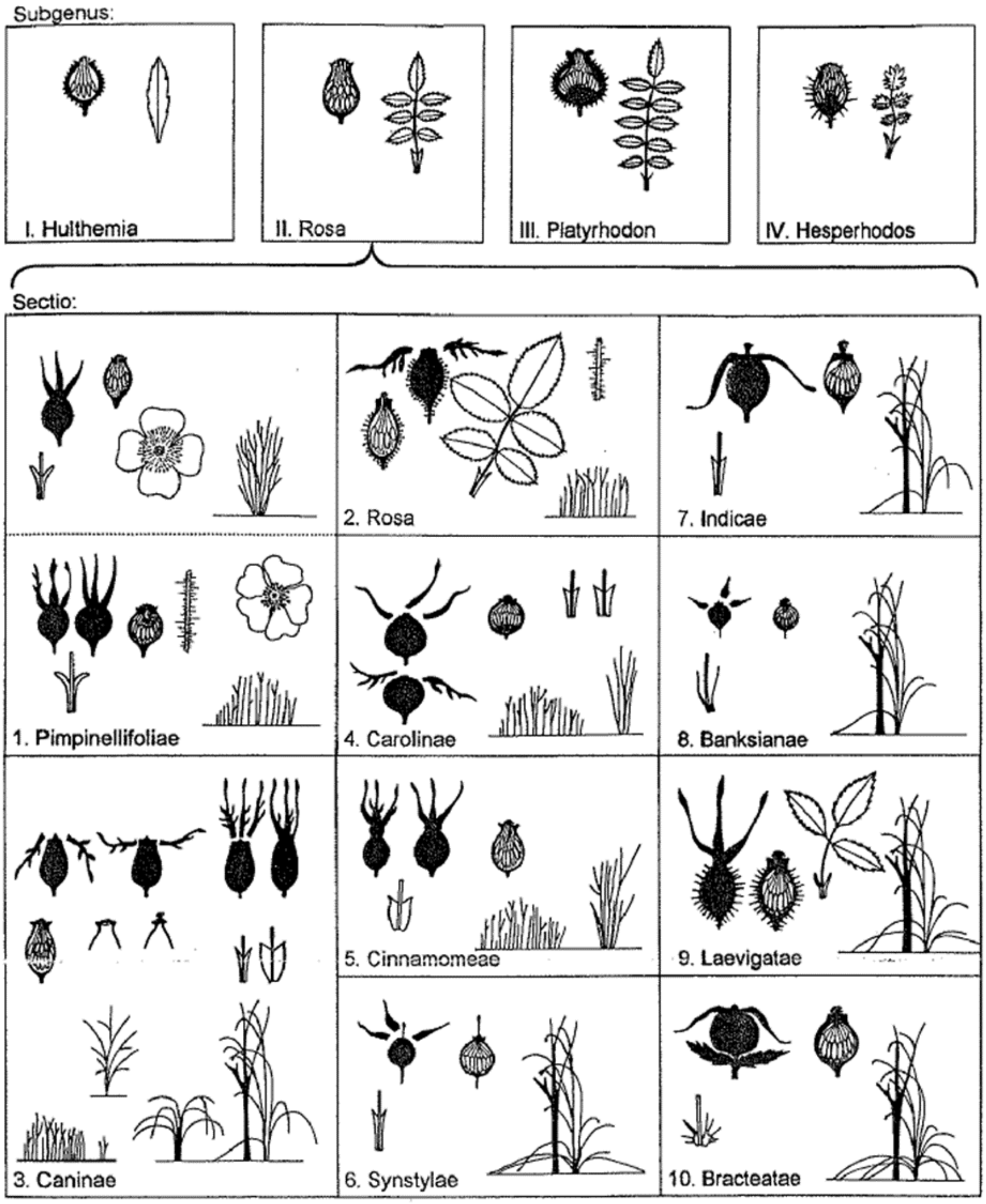


Figure 14: Principal traits associated with the classification of *Rosa* into subgenera and sections. Extracted from Henker (2000).

tical to oblong, serrated towards the apex, most often pubescent and very slightly spinous on the reverse, with serrulate margins. The flowers are small, odourless, solitary, with a diameter of 2 to 2.5 cm on a pedicel of 1 to 1.5 cm; absence of bracts; lanceolate sepals; golden yellow petals with a purple or brown basal spot; yellow stamens turning quickly to purple yellow. Flowering in early summer. The fruits are spherical, purple brown, almost black, very thorny, persistent sepals. *Rosa persica* is not a hardy species, fearing severe frosts but tolerant to drought and thriving in stony, well drained soil and sunny places.

Sub-genus *Platyrhodon* (Hurst) Rehd.

It is exclusively dedicated to *Rosa praelucens*, the decaploid rose, and *R. roxburghii* Tratt. and its varieties. *R. praelucens* is an endangered species. These plants originate from the Far East and are distinguished by their peeling bark and large thorny hips, not very durable, which appear after flowering and which have earned *R. roxburghii* the nickname of chestnut rose.

Sub-genus *Hesperhodos* Cockerell

It contains two species, *R. minutifolia* Engelm. and *R. stellata* Woot., originating from the southern part of North America (California, Baja California). They are low bushes, with small pinnate leaves composed of 3 leaflets, sometimes 5 and more rarely 7, bracts are absent and flowers, most often solitary, are pink, white or purple. The fruits are covered with stiff hairs.

Sub-genus *Rosa*

It includes nearly 150 species and is subdivided into ten sections.

Section *Banksianae* Lindl.

Native to China, the two climbing rose species (*Rosa banksiae*, *R. cymosa*) in this section have many yellowish or white flowers arranged in umbels or compound corymbs. Their long glabrous stems or with a few hooked prickles can reach up to 10 m long; their evergreen pinnate leaves are composed of 3 to 7 leaflets, the petiole being provided with detached and deciduous stipules. Small deciduous bracts, curved and deciduous sepals are other distinctive features of these two species. These plants are sensitive to frost.

Section *Bracteatae* Thory

Two Asian species of climbing roses fall under this section, one from India (*Rosa clinophylla*) and the other from the warm regions of southern China (*R. bracteata*). The shoots carry pairs of curved prickles implanted under the evergreen and shiny leaves. The leaves are pinnate and composed of (5) 7 to 9 (11) leaflets and have adneous stipules. The flowers, white or ivory white, solitary, occasionally in few-flowered bouquets, with a tomentous receptacle and large bracts, characterize these roses, which do not tolerate extreme cold.

Section *Caninae* DC.

This section includes about thirty species of wild roses from Asia, Europe and North Africa, often armed with strong hooked prickles, common along roadsides, in country hedges and wastelands. The leaves are pinnate and composed of 5 to 7 leaflets, sometimes 9, pubescent or glandular. They are generally white or pale pink flowering roses, devoid of bracts and most often solitary, with sepals falling after flowering. Interspecific crosses are frequent.

Section *Carolinae* Crép.

This section includes six species of roses native to North America. These are low bushes with thin stems on which many curved or straight prickles point, implanted under the leaf nodes. The leaves, often bright, are composed of 7 to 9 leaflets. After the flowers, solitary or in few-flowered bouquets, appear the fruits, more or less globular from where the sepals, deployed after flowering, quickly detach.

Section *Chinenses* DC. (syn. *Indicae* Thory)

Three species from China and Myanmar are included in this section. They are erect or climbing shrubs with hooked stems. The leaves, composed of 3 to 5 (7) leaflets, have narrow stipules, adnate to the petiole, with tapered and divergent auricles. Usually grouped in bouquets, the flowers have white, pink, yellow or red petals and sepals reflected after flowering, deciduous before the maturity of hips. Free styles are half as long as stamens. The species in this section, imported into Europe at the end of the 18th century, are at the origin of recurrent blooming or continuous flowering varieties.

Section *Rosa* (syn. *Cinnamomeae* DC)

This section includes about 50 species from Asia (36 from China), Europe and North America. “Cinnamon” roses, with pink, red, lilac and magenta flowers, more rarely white, are erect, tall, often suckering bushes, with stems generally armed with straight or hooked prickles, with the exception of floral stems which are most often glandulo-pubescent. The leaves, sometimes persistent, are composed of 5 to 11 (15) leaflets preceded by adnate stipules, with dilated and divergent auricles. The flowers, rarely solitary, have erect sepals, usually persistent after flowering. The peduncles carry more or less dilated bracts.

Section *Gallicanae* DC

This section contains several species and ancient old hybrids from Anatolia and Europe. They are erect, low bushes with stems armed with curved prickles often mixed with glandular bristles. The leaves, quite firm and provided with adneous stipules, are generally composed of 5 leaflets, more rarely 3. Carried on long floral stems, the flowers, most often solitary, have a variety of colors ranging from white to pink and purple (yellow is absent). The flowers are sometimes variegated, striated with white stripes. The sepals, reflected after flowering, persist on the fruits until their maturity but fall shortly before it. The stems of multiflowered inflorescences often have small, narrow bracts. Hybrids are numerous and appeared in Europe at the beginning of the 15th century.

Section *Laevigatae* Thory

Only one species, originating from China, is included in this section: *Rosa laevigata* Michx. It is a climbing or trailing shrub, armed with curved prickles often mixed with bristles. The leaves, persistent and usually composed of three tough and bright leaflets, have free or nearly adnate stipules that detach easily. Carried on a spinous pedoncule, the large white flowers, solitary and with many stamens, have sepals that persist after flowering and stand on the hip. Bracts are absent. This species is not very resistant to extreme cold.

Section *Pimpinellifoliae* DC.

The species gathered in this section are low bushes with white, pink, bright yellow or purple flowers, originating from Asia and Europe and rarely exceed 3 m in height. The stems, erect, are covered with bristles or straight prickles of varying size; *Pimpinellifoliae* species carry small leaves composed of 7 to 9 leaflets, rarely more than 15, resembling those of the burnet (*Sanguisorba minor*, FR: Pimprenelle), with narrow stipules, long adnate, dilated and divergent auricles. Bracts are absent. The flowers, solitary, have simple sepals and usually stand on the hips upon which they remain attached after maturation. Some rare species have flowers with only 4 petals and 4 sepals.

Section *Synstylae* DC.

This section encompasses about twenty species with white, pink or purple flowers, present throughout the distribution area of the genus, but more especially in Asia, from Korea to Turkey. These plants are among the ancestors of most of the climbing rose varieties produced by horticulture. They grow in large shoots, covered with curved or hooked stings and leaves composed of 5 to 7 (9) leaflets, with stipules that are long adenate. Bracts are usually absent. Their flowers, numerous and arranged in corymbs or small 3-flowered bouquets, have reflected sepals that quickly fall after the fruit ripening. The styles are stick together in a small column.

1.3.5 Classification in the light of molecular data

Overview of the relevant methods

During the twentieth century, technical advances in science shed new light on the classification of roses, first with the contribution of cytology (chromosome counts in metaphase) and then with the analysis of DNA markers and sequences. These new techniques provided more features to group species into subgenera and sections. Molecular sequences correspond to specific arrangements of a 4-base alphabet in the case of DNA/RNA and a 20-amino acid alphabet for protein. DNA

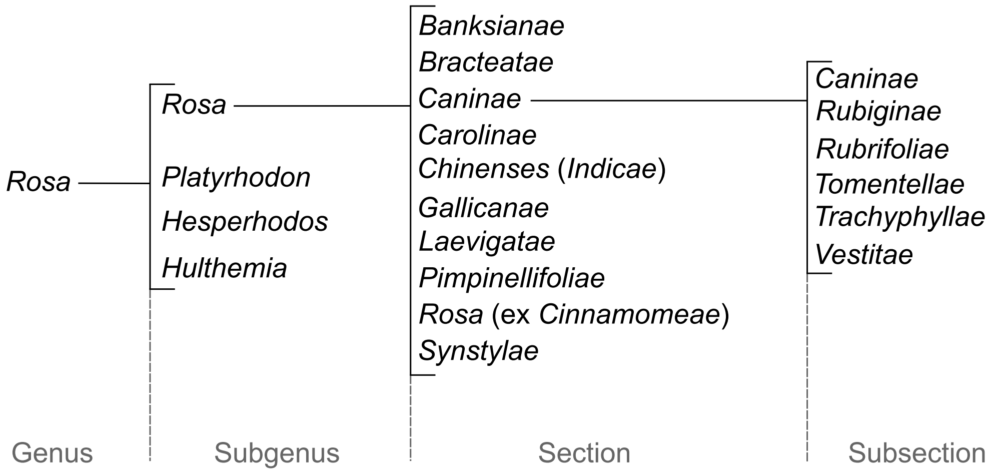


Figure 15: Commonly used classification of the genus *Rosa*. The present classification mainly corresponds to Rehder (1940) with slight modification by Wissemann (2003a). This classification relies only on morphological characters and includes updates about subgenera and section names.

sequences generally contain more variations, especially in non-coding regions (Igea et al., 2010), and are therefore more adapted to study close species relationships, while protein sequences are more conserved and easier to align when more divergent taxa are studied (Michu, 2008). The haploid nuclear genome sequence of *R.* 'Old Blush' encompasses a bit more than 500 Mbp (Raymond et al., 2018; Hibrand Saint-Oyant et al., 2018) that provide an almost inexhaustible number of characters which can be compared between species for phylogenetic inferences. Molecular sequences are generally aligned prior to infer phylogenetic relationships. The resulting alignments enable to compare each position of the sequences between a set of taxa and identify similarities or differences that help distinguish groups of taxa that share derived traits which is clear evidence of close relationships. Not all sequence variations are informative. Indeed, most phylogenetic studies based their selection of phylogenetic markers on the amount of variations rather than on the informativeness of such variations. In fact, only (syn)apomorphies (ie shared derived characters obtained from a common ancestor (Figure 16)) are useful to group species into monophyletic clades. However, when fast-evolving sequences are used, homoplasies (ie same derived character that appear spontaneously in different organism without sharing a common ancestor bearing that trait (Figure 16)) are likely to occur and may overwrite (syn)apomorphies and lead to biased relationships.

The classic output of phylogenetic inference is a tree, that shows the different relationships between species. Once molecular sequences are aligned, there are two sorts of phylogenetic tree building methods that can be used to study species phylogenetic relationships: distance-based methods and character-based methods (Yang and Rannala, 2012). Distance-based methods convert sequence alignments into a matrix of pairwise distances between sequences and use it to compute branch lengths and tree structure. Neighbour joining method (NJ) (Saitou and Nei, 1987) and the

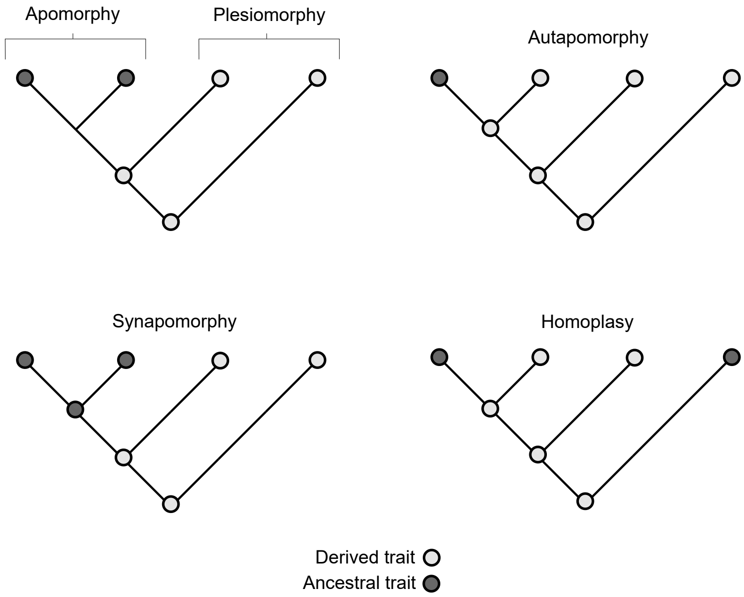


Figure 16: A schematic representation of the possible configurations for ancestral and derived traits. Apomorphy is a shared trait between two or more species and that sets the clade apart from other clades. If the derived trait also belongs to a common ancestor, the term synapomorphy can be used in place of apomorphy. Autapomorphy corresponds to a derived trait that is unique to one taxon and is useless in phylogenetics. Plesiomorphy corresponds to a shared ancestral trait among two or more taxa. Homoplasy is a derived trait that appeared independently in different taxa. Only (syn)apomorphy and plesiomorphy are informative for phylogenetics. Adapted from Page and Holmes (2009).

unweighted pair group method with arithmetic means (UPGMA) (Sokal and Sneath, 1963) are two distance-based methods that are commonly used to infer relationships between species.

Character-based methods aim to optimize one or several parameters (for example number of transformations in maximum-parsimony methods) on all possible relationships between taxa. To do so each site is analyzed independently, using an explicit evolutionary model. Such approaches are assumed to be best suited when mutation changes are not homogeneous among lineages. Unlike character-based approaches, distance-based methods merge all the information provided by individual sites in a single value, mathematically computed, and construct a tree step-by-step. Such distance-based methods may be valuable when mutation rates are homogeneous in all branches of the tree, but going back to the relevant character explaining a particular relationship is then more complicated. There exist many evolutionary models of DNA/protein sequences. They generally assumed that evolution is independent among lineages and memory-less, that the sites evolve independently and in the same manner across individuals, models should be time-reversible (Liò and Goldman, 1998) and gaps are often treated as unknown data. Each model of evolution is associated with a bunch of parameters that are estimated from the alignment being analyzed. Evolution models vary according to parameters that correspond to the proportion of variable sites, the homogeneity (or not) of mutation rates between bases/amino acids, with sometimes a distinction according to the nature of the bases/amino acids. For DNA, the simplest model corresponds to the Jukes Cantor (JC69) (Jukes and Cantor, 1969) with 0 parameter and the more complex is the 8-parameter GTR model (Tavaré, 1986) (Figure 17).

Character-based methods include three main methods: maximum parsimony, maximum likelihood and Bayesian inference. Maximum parsimony seeks the tree with the minimum number of changes needed to convert one sequence to another along the tree, assuming that the most likely scenario corresponds to the one with the minimum number of events along the tree (Yang and Rannala, 2012). Maximum likelihood and Bayesian inference are expanded model-based methods. Maximum likelihood searches the tree that maximizes the probability of observing the data given that tree while Bayesian inference looks for a subset of plausible trees among a set of trees obtained after several generations of convergent parameter estimations given prior on DNA model evolution and parameter (Yang and Rannala, 2012). In any case, validation methods are generally used to assess the support of each branch of the tree (Alfaro et al., 2003). Bootstrapping correspond to a random resampling with replacement of positions (columns) in sequence alignments to evaluate the robustness of each branch (Felsenstein, 1985) and is commonly used in maximum parsimony and maximum likelihood phylogenetics. Bootstrap supports above 70% are generally considered to show supported relationships (Hillis and Bull, 1993). In Bayesian inferences, posterior probability corresponds to the number of times a specific branch is found in a set of trees sampled after a certain number of generations of convergent parameter estimations (Lewis, 2001; Huelsenbeck and Ronquist, 2001). Posterior probabilities above 95% are clear evidence of well supported relationships. Unlike distance-based methods, character-based methods relying on complex evolutionary model with thorough validation procedures are penalized due to the associated computational bur-

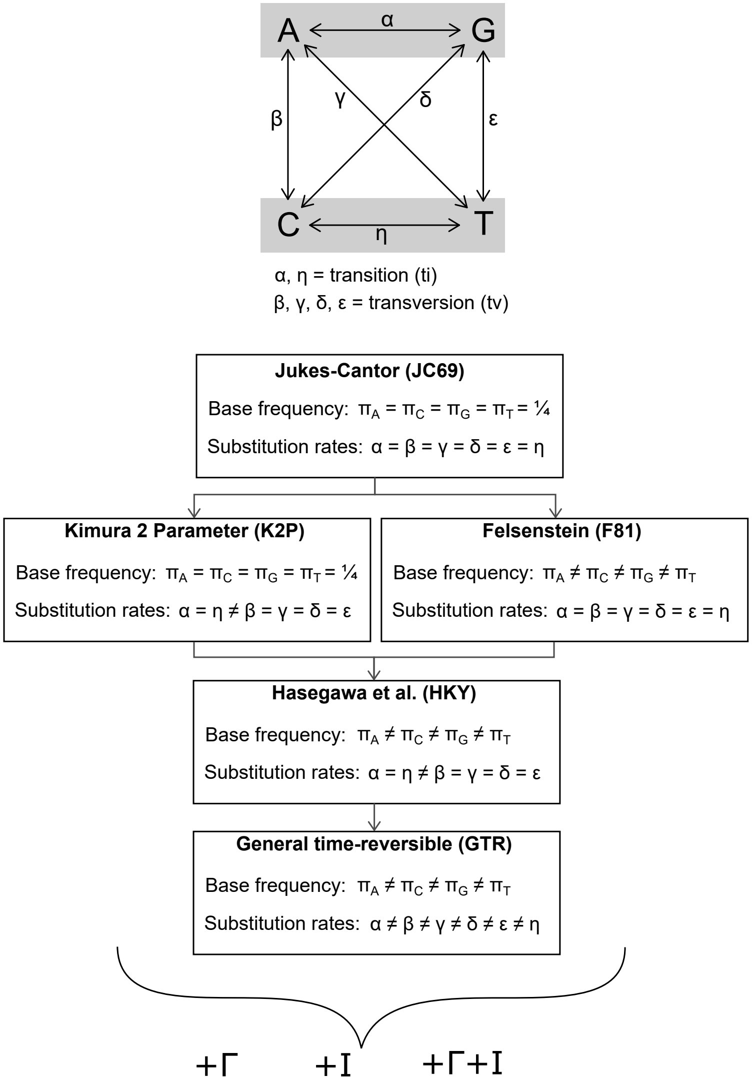

A
G
C
T
α
β
γ
δ
ε
η
α, η = transition (ti)
β, γ, δ, ε = transversion (tv)
Jukes-Cantor (JC69)
Base frequency: π_A = π_C = π_G = π_T = ¼
Substitution rates: α = β = γ = δ = ε = η
Kimura 2 Parameter (K2P)
Base frequency: π_A = π_C = π_G = π_T = ¼
Substitution rates: α = η ≠ β = γ = δ = ε
Felsenstein (F81)
Base frequency: π_A ≠ π_C ≠ π_G ≠ π_T
Substitution rates: α = β = γ = δ = ε = η
Hasegawa et al. (HKY)
Base frequency: π_A ≠ π_C ≠ π_G ≠ π_T
Substitution rates: α = η ≠ β = γ = δ = ε
General time-reversible (GTR)
Base frequency: π_A ≠ π_C ≠ π_G ≠ π_T
Substitution rates: α ≠ β ≠ γ ≠ δ ≠ ε ≠ η
+Γ
+I
+Γ+I

Figure 17: Nested models of DNA evolution. Each model relies on two kinds of parameters: base frequencies and substitution rates. These parameters can be fixed or estimated empirically from the sequence alignment. The simplest model (JC69) implies that all base frequencies and substitution rates are equal while the most complex model (GTR) allows for different base frequencies and substitution rates. All models can be completed with additional parameters (Γ and I) to add significantly more realism to the model chosen. Γ models substitution rate heterogeneity over alignment sites using a gamma distribution (Proportion of sites = f(substitution rate)) with one shape parameter. I enables to consider invariant sites in the modeling of rate heterogeneity.

den. However, they are generally preferred since their evolutionary models have more biological meaning than distance-based models. Moreover, in the character-based approach, it is possible to find those characters that explain a particular relationship, whereas this is not explicitly done in distance-based approaches since the information is aggregated.

Applications of cytology and molecular sequencing to *Rosa* phylogenies

Hurst (1925) was one of the first scientists to bring valuable information about chromosome counts and morphology in the genus *Rosa* since he provided new evidence about genomes relationships in the light of morphological traits. He demonstrated the existence of seven basic chromosome forming a septet and identified five different morphological groups that could be consistent with his chromosomes counts. He named them from A to E and distinguished regular diploid AA to EE from polyploids that can be regular (ie with bivalent septets) or irregular (mix of bivalent and univalent septets). He also treated wild polyploids as allopolyploids (differential polyploid species) and provided some perspectives on their putative origin using combinations of the five septets in relation to morphological observations. However, Hurst's classification was questioned notably because it is based on the assumption that species belonging to a same septet group are interfertile, which has been proven not to always be the case by further hybridization studies (Lewis and Bayse, 1961).

Later, in 1990s, molecular sequences provided numerous new characters to study *Rosa* species relationships and proved to be useful to both delimit species and study their evolutionary history. Numerous attempts were done to end up with a comprehensive evolutionary history of *Rosa* using a wide range of molecular markers: Rapidly Amplified Polymorphic DNA (RAPD) (Millan et al., 1996; Jan et al., 1999; Atienza et al., 2005; Atif Riaz, 2011), Amplified Fragment Length Polymorphisms (AFLP) (Koopman et al., 2008), microsatellite markers (SSRs) (Scariot et al., 2006; Zhang et al., 2013), DNA sequences from plastid gene interspacers (Matsumoto et al., 1998; Wissemann and Ritz, 2005; Bruneau et al., 2007; Qiu et al., 2013; Fougère-Danezan et al., 2015; Zhu et al., 2015), nuclear ITS (Wu et al., 2001; Wissemann and Ritz, 2005; Qiu et al., 2012, 2013; Zhu et al., 2015) and low copy nuclear gene of *GAPDH* (Joly et al., 2006b; Meng et al., 2011; Fougère-Danezan et al., 2015; Zhu et al., 2015).

Virtually all categories of phylogenetic tree building methods have been applied to study the evolutionary history of *Rosa*, either alone or combined, from distance-based methods (UPGMA (Jan et al., 1999; Scariot et al., 2006; Koopman et al., 2008)) to character-based methods (Maximum parsimony (Matsumoto et al., 1998; Wu et al., 2001; Bruneau et al., 2007; Koopman et al., 2008; Qiu et al., 2012, 2013; Zhu et al., 2015)), Maximum likelihood (Zhu et al., 2015; Fougère-Danezan et al., 2015), Bayesian inference (Wissemann and Ritz, 2005; Koopman et al., 2008; Meng et al., 2011; Fougère-Danezan et al., 2015; Zhu et al., 2015)).

The morphological sections of Rehder are generally not found monophyletic in molecular analysis (see Figure 18 for definitions of monophyly vs paraphyly, Figure 19).

Especially, the monophyly of *R.* subg. *Rosa* is not confirmed and the other subgenera are

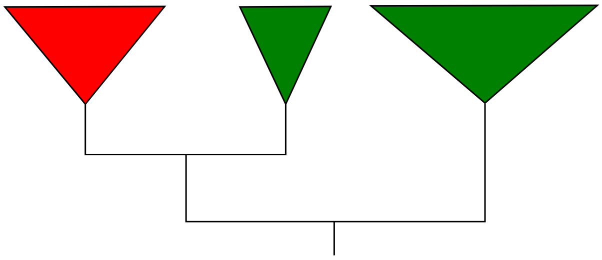


Figure 18: A schematic representation of monophyly and paraphyly. Red species form a monophyletic group since they descend from a common ancestor. Green species are split into two distinct clades with different common ancestors so they form a paraphyletic group.

often embedded within the subgenus *Rosa*, therefore authors suggest to treat subg. *Platyrhodon*, *Hulthemia* and *Hesperhodos* at the sectional level (Wissemann and Ritz, 2005; Fougère-Danezan et al., 2015). In the last and most complete phylogenetic analysis of *Rosa*, Fougère-Danezan et al. (2015) identified two main clades that split the genus, namely a *Cinnamomeae*-like clade and a *Synstylae*-like clade (see Figure 2 in Fougère-Danezan et al. (2015)). The *Cinnamomeae* clade gathers species belonging not only to *R.* sect. *Cinnamomeae* but also to *R.* sect. *Carolinae* and some species from the *Pimpinellifoliae* section. They mostly correspond to Asian and North American species, although some are also found in Europe (*R. majalis*, *R. pendulina*, *R. spinosissima*). The *Synstylae*-like clade encompasses species from *R.* sect. *Syntylae*, *Chinenses*, *Caninae* and *Gallicanae* which usually thrive in Asia and Europe (*R. setigera* (*Synstylae*) is native to America). The mono- bi- specific sections *Banksianae*, *Bracteatae*, *Laevigatae* tend to be closer to the *Synstylae*-like clade. The position of the section *Pimpinellifoliae* is doubtful because some consectional species spread across different clades of the phylogenetic trees. Nevertheless, the *Pimpinellifoliae* section seem to be closer to the *Cinnamomeae*-like clade.

Molecular phylogenetics of *Rosa* greatly improved knowledge on species relationships because they identified clades that do not always correspond to the morphological classification. However, phylogenetic trees of Rosa generally lack support, especially for deep branches. This means the relationships between groups, sometimes corresponding to section and subgenera, are not well supported and it is thus difficult to infer a general evolutionary history of the genus. There are many underlying technical and biological reasons that could explain why most of deep branches are not well supported.

First, the sequences and markers that were used generally concentrate inherent shortcomings. As for AFLP and RAPD, they might evolve too rapidly and create homoplasy that therefore hamper accurate study of ancient speciation events (García-Pereira et al., 2010). Regarding DNA sequences, authors generally resorted to chloroplast sequences which are only inherited from one parent in plant, generally the mother in angiosperms (Reboud and Zeyl, 1994), thus providing a uniparental orientated view of the evolution. Such bias may not be problematic at high taxonomic ranks since differences between maternal and paternal lineages may have been blurred by a

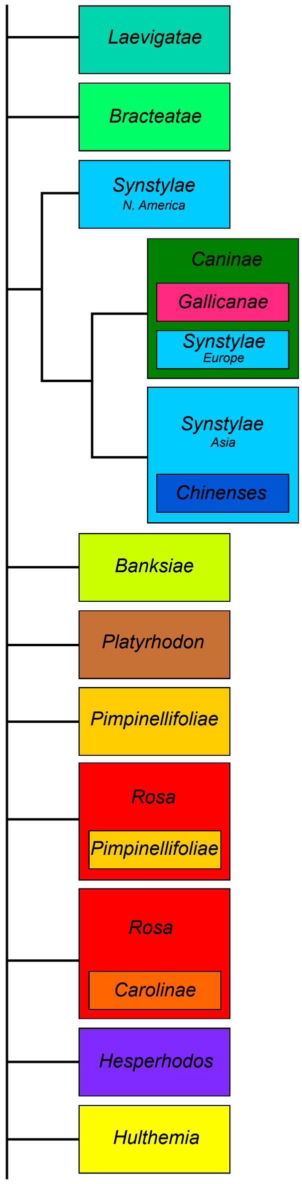


Figure 19: Schematic representation of the phylogenetic relationships between subgenera and sections in *Rosa* according to Fougère-Danezan et al. (2015). Branches with bootstrap support <70% were collapsed.

very large number of generations. However, as soon as differences between maternal and paternal lineages are recent, or somehow fixed in the genome (through duplications, or asymmetrical transmissions in gametes), reticulated-like events cannot be appreciated using plastid sequences alone (Sang, 2002). This may be the most important criticism that can be done toward the use of plastid sequences alone to reconstruct phylogenetic relationships in *Rosa* since (1) both hybridization and polyploidy are major driving forces, especially at recent times, and (2) most speciations in *Rosa* are recent (<5 MY) (Fougère-Danezan et al., 2015). The paternal contribution is totally ignored and reticulated pattern are therefore omitted. In addition, plastid sequences are known to generally be less variable than nuclear genomes and thus their phylogenetic informativeness can be less appropriate to study close species relationships (Sang, 2002). When nuclear DNA sequences were used, in combination with plastid sequences or not, authors again targeted ubiquitous sequences, namely ITS or *GAPDH*. As for ITS, these sequences exist in many copies in plant genomes and mostly correspond to a paralogous gene family (Naumann et al., 2011). This means that phylogenetic studies involving ITS may include sequences which (1) do not originate from a speciation event but rather from duplications or (2) that have undergone concerted evolution (Wendel et al., 1995; Eickbush and Eickbush, 2007; Naidoo et al., 2013). In concerted evolution, nuclear ribosomal ITS copies do not evolve independently and therefore go against classical evolutionary expectations. However, Wissemann (2003b) reported non-concerted evolution of nrITS in *R.* sect. *Caninae* and *Gallicanae* (ex *Rosa*), which could be an evidence for recent and rapid radiations of these sections. Renny-Byfield et al. (2011) even reported that paternally derived ribosomal DNA was lost after allopolyploidization in tobacco (*Nicotiana tabacum*), thus questioning the utility of rDNA and associated nrITS in the study of reticulated patterns in plants. The resulting evolutionary history may therefore be biased and may provide inaccurate phylogenetic relationships (Poczai and Hyvönen, 2010). Concerning *GAPDH*, while it has been found the most variable ubiquitous low copy gene in Rosa (Joly et al., 2006b), no verification could have been made to validate the specificity of the primer pair since there was no genome sequence at that time. In addition, *GAPDH* sequences are not that variable across close related *Rosa* species (Joly et al., 2006a) and their was at that time no conceptual index to assess their phylogenetic informativeness and see their ability to resolve different epochs of the evolutionary history of *Rosa*. Therefore, Zhu et al. (2015) concluded that phylogenetic trees are less resolved with nuclear *GAPDH* than with plastid sequences. Zhu et al. (2015) suggested two main possibilities to explain conflicting gene tree: hybridization and incomplete lineage sorting (ILS). ILS is a situation that occurs when a gene tree differs from the species tree, therefore producing a discordant tree. This is due to the fact that some alleles did not coalesce (looking backward in time) into an ancestral allele until time deeper than speciation events. For instance, in Figure 20, the ancestral population of species A, B and C had three alleles marked in blue, green and yellow respectively. Allele B in the lineage leading to species B failed to coalesce in the recent T1 speciation event but coalesce with allele C before T2. This ILS may result in conflicting topologies between gene tree and species tree. ILS is likely to occur in recent taxonomic groups, such as species inside genus because too few time generally separate the last speciation event from the present. In *Rosa*, ILS seems likely to occur since some lineage radiated only few million years ago. However, Zhu et al. (2015) suggested that hybridization is the major factor explaining conflicts between gene trees and species tree in *Rosa*. However, we argue that forthcoming *Rosa* studies should resort to sequences and tree building method able to consider both hybridization and ILS.

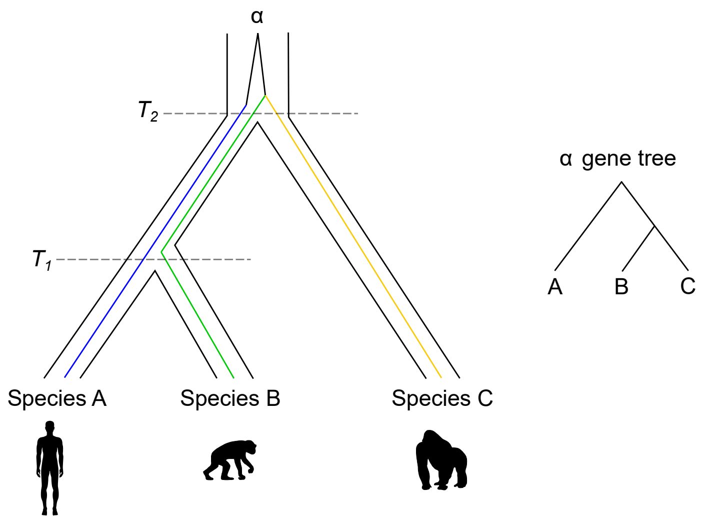


Figure 20: A schematic representation of incomplete lineage sorting (ILS). T1 and T2 indicate times of speciation events. The black frame corresponds to the species tree topology showing the phylogenetic relationships between 3 species: A (Human), B (Chimpanzee) and C (Gorilla). The gene α has three different allele sequences (blue, green and yellow) corresponding to the three respective species. Allele B in the lineage leading to species B failed to coalesce in the recent T1 speciation event but coalesce with allele C before the previous T2 speciation event. The resulting gene tree of α (right) incorrectly shows that species B (Chimpanzee) is closer to species C (Gorilla) than to species A (Human), this is called ILS.

Second, *Rosa* phylogenetic analyses were mainly based on few markers and sequences. Indeed, analyses usually resort to two or three loci for an overall concatenated matrix of less than 1500 bp which might be insufficient to obtain high support along branches of the phylogenetic tree despite some fast-evolving regions (Wortley et al., 2005). Moreover, the targeted sequences were generally identified thanks to previous studies in other taxonomic groups and were not assessed for their ability to solve both ancient and recent speciations in *Rosa* since such approaches did not exist. The presence of informative indels was not always considered in the last *Rosa* phylogenies (Fougère-Danezan et al., 2015) despite they may bring informative variations to help distinguish between lineages.

To sum up, phylogenetic relationships among *Rosa* were investigated using various markers and methods but often led to contradictory results, mostly because branches of the phylogenetic tree lacked supports. Several shortcomings can be attributed to the markers/sequences and methods that have been used up to now, providing much room for further improvements of the *Rosa* phylogeny using optimized sequences and approaches.

1.4 Perspectives on the classification of the genus *Rosa*

1.4.1 Accessing wild roses through the world

Collecting *Rosa* specimens directly from the wild, where they grow naturally, seems a priori the best way to access quality material. However, there are some obstacles to collect rosebush samples in their natural habitats. First, it required significant skills in botany to be able to distinguish between species, which can be quite difficult for very close related taxa (Kellner et al., 2014). Second, wild roses are found throughout the Northern Hemisphere and sampling the genus *Rosa* itself would require many expeditions and may take decades to achieve. Third, some *Rosa* species are (locally) listed as endangered species and benefit from special protection policies that limit free collections of samples in the wild. This is notably the case for *R. gallica* in France, *R. minutifolia* in USA and *R. praelucens* in China. In addition, recent international protocols, such as the Convention on Biological Diversity (Rio de Janeiro 1992) and the Nagoya Protocol on Access and Benefit Sharing (Nagoya 2010), supervise the collection of wild material and may therefore limit their use and sharing, even for public research and education (Deplazes-Zemp et al., 2018; Prathapan et al., 2018). Consequently, the choice to prospect samples in the wild must be considered against other methods to access material, in particular well-documented ex-situ collections.

Thanks to the general interest in roses, numerous rose germplasm resources exist through the world. However, due to historical and sociological constraints, associated with the vast range of climate conditions under which wild roses can grow, the germplasm is scattered across many different collections. Even the few rose gardens fully dedicated to botanic/wild roses do not gather all species of *Rosa* individually. Therefore, there is no reference collection that encompasses the whole genus diversity. Access to each wild rose species has to be treated on a case-by-case basis. Hopefully, several wild roses are part of botanical heritages and could have been preserved locally in botanic gardens at the initiative of local environmental policies, in relation to their missions of conservation, research or education. Therefore, accessing accurate germplasm may be better achieved when asking sample material to local botanic gardens. Furthermore, most rose gardens maintain some botanic accessions derived from specimens collected in the wild along with modern cultivars. Unlike botanic gardens that are geared towards conservation of wild material, the purpose of rose gardens is to display the diversity present in cultivated material. Therefore, passport data relating to the origin of botanic specimens in rose gardens might be inaccurate or incomplete. In the quest for quality samples, it is worth mentioning the effort of centralization carried out by open access databases such as PlantSearch which enable to contact worldwide botanic gardens that specifically grow the requested species. This type of initiative has to be supported since it bridges the gap between conservation, research and education on a global scale. It enables win-win opportunities to better understand plant evolution and biology as well as promoting species conservation. As for *Rosa*, it may be interesting to collect several specimens from different botanic garden to have a better view of the intraspecific genetic diversity and prevent errors due to mislabeling. The minimum passport data required to describe a specimen are quite homogeneous across worldwide botanic collections and some botanic gardens even resort to experts for regular inspections of their collections to detect misidentified accessions. In addition to these preventive measures, rose collectors should resort to specific DNA barcoding to test the correct assignment of their specimens. However, no DNA reference barcodes and general procedures exist up to now to clearly distinguish between *Rosa* species and discard non-natural hybrids, although some work was carried out in this direction (Schori and Showalter, 2011).

As a last option, wild specimen might be accessed through ex-situ collections of dead (dried plant fragments in herbaria) or dormant (seeds) material. For rare species, herbaria specimens offer a nice alternative to living material for which sample access may be limited. The largest rose herbaria in the world may correspond to the collection present at The Herbarium of the Botanic Garden Meise, Belgium, with 50,000 rose specimens, both wild and cultivated. A large part of this rose herbarium corresponds to Crépin's collection (40,000 specimens). However, the conditions under which specimens were dried and preserved greatly influence the yield of DNA extractions. A pilot study on the Crepin's herbaria suggests that a substantial number of samples could be processed for DNA extractions and would lead sufficient amount of quality DNA molecules for amplifications of nrITS and plastid loci (Stoffelen et al., 2018). Given the ever-growing accuracy of DNA technologies, these results are promising for future applications of Next-Generation Sequencing techniques on such material. For seed banks however, there are limitations to using rose seeds for phylogenetic applications because (1) rose germination implies a cumbersome work (stratification to break the embryonic dormancy, embryo rescue, in-vitro propagation) and (2) roses hybridize easily so unless seeds come from a wild population, the use of rose seedlings for phylogenetics may provide uncertain results.

1.4.2 The advent of Next-Generation Sequencing techniques

Rosa classification has been extensively studied through the lens of morphological markers. While they proved to be useful to distinguish between the main sections, they are of limited interest for further investigating phylogenetic relationships between intrasectional species. This is mainly because too few reliable morphological characters can be used to distinguish one species from its close related counterparts. From the 1990s onwards, molecular sequences obtained with Sanger sequencing broadened the possibility to model the evolutionary history of *Rosa*. The resulting phylogenetic trees further served as a basis to revise the genus *Rosa* classification, but the number of available molecular sequences was limited and so were the conclusions. Nowadays, Next-Generation Sequencing techniques outshine all the shortcomings of Sanger sequencing in terms of scalability and virtually provide access to any variations in genomes. Such variations can be wisely used to infer phylogenetic relationships between organisms, considering whole genome sequences. The recent gain in sequencing knowledge led to tens of *Rosa* whole genome shotgun datasets available on public gene databases, as well as high-quality reference genomes (Raymond et al., 2018; Hibrand Saint-Oyant et al., 2018). In the following paragraphs, we review the main DNA technologies that are currently used to obtain molecular sequences.

Sanger sequencing is currently the most used technique to rapidly and cheaply obtain the sequence of a nucleic acid. It is based on the selective incorporation of chain-terminating dideoxynucleotides (ddNTPs) by DNA polymerase during in vitro replication (Sanger et al., 1977). The experiment relies on four different reaction mixes and has been further optimized to be used with fluorescent ddNTPs (Smith et al., 1985, 1986). All four mixes include the double stranded DNA containing the region to sequence, a primer flanking the region of interest to initiate the elongation, a DNA polymerase and the four dNTPs (A, C, G, T). However, each mix contains one of the four ddNTPs which is marked by fluorescence (ddNTP*). During the elongation process, the elongation stops if a ddNTP* is included. This results in many fragments of different length with each a ddNTP* in its 3' end. The different fragments can be distinguish using a capillary electrophoresis at a precision of 1 bp. The resulting gel is read gradually by an optical laser able to recognize the wavelength emitted by the 3' ddNTP* of each fragment thus rendering the full DNA sequence. Sanger sequencing is still the preferred way to sequence DNA in small scale projects (Bibault and Tinhofer, 2017), however it is less cost-effective when dealing with many sequences across many samples (Metzker, 2010). Virtually all gene sequences obtained up to now for *Rosa* phylogenetics were sequenced using the Sanger sequencing technique.

Next-generation high-throughput DNA sequencing techniques (NGS) are now extensively used to investigate genome scale variations (Koboldt et al., 2013; Wuyts and Segata, 2019) and are each year more affordable (Wetterstrand, 2019). They allow to sequence DNA molecules at depth and coverage out of range of Sanger sequencing. For large scale genomic exploration, they represent efficient, accurate and cost effective alternatives to Sanger sequencing. NGS experiments usually require the construction of NGS library that consist of similar size DNA fragments which 5' and 3' ends are completed with sequencing adapters. The following steps of sequencing depend on the method (Reuter et al., 2015). To simplify, each method requires a kind of cell which differs in its composition and structure from one technique to another. There are mainly to type of sequencing techniques depending on the length of the output reads. (1) Short read sequencing techniques that are largely dominated by Illumina technologies. Illumina technology uses a glass flow cell that contains millions of oligonucleotides on which sheared DNA can bind and be gradually amplified thus forming clusters. The read sequencing then relies on the detection by optical laser of fluorescent dNTPs incorporations by DNA polymerases at the cluster level. (2) Long read sequencing techniques are also based on cell but they contain either nanowells or nanopores. The former corresponds to zero-mode waveguides (ZMWs) that resemble wells able to detect the incorporation of fluorescent dNTPs during the elongation of one single DNA molecule by a DNA polymerase (Levene et al., 2003; Eid et al., 2009; Reuter et al., 2015). The latter refer to small holes made of proteins able to conduct one DNA molecule from one side of the cell to the other. The nanopores are immerged in an electrolytic solution so that an electric current can be observed through the nanopore. This electric current changes depending on the molecule present in the nanopore and is different for each of the four nucleotides (Branton et al., 2008). Therefore, long DNA molecules can be read from start to end and limit the number of reads to reassemble for whole genome sequence reconstruction. Currently, both short and long read sequencing techniques provides pros and cons (Goodwin et al., 2016). On one side, short read sequencing is more accurate and somewhat faster for similar yields but it generates highly fragmented genomes which might be difficult to assemble. On the other side, long read sequencing is less accurate but provides longer reads giving unprecedented insights in genome structures. Long read sequencing could especially be an effective way to retrieve homologous chromosome fragments and would considerably ease the steps of assembly. Both short and long read sequencing technologies are usually combined to achieve high quality genome sequencing. The DNA sequencing market is constantly evolving as new technologies are developed or improved. These techniques are increasingly more efficient and less expensive and there is no doubt that the number of sequencing projects will increase in the forthcoming years (Kumar et al., 2019).

1.4.3 From phylogenetics to phylogenomics

For quite a long time, molecular phylogenetics has been dominated by few ubiquitous DNA sequences. In plants, sequences from the chloroplast genome were extensively used since they were conserved enough to be amplified in non-model organism and variable enough to still distinguish between most of species (Shaw et al., 2005). In addition, plastid genome is haploid which avoid to deal with several homologous sequences. This greatly facilitates rapid evolutionary studies of green plants (Ruhfel et al., 2014), especially for complex taxa involving polyploid specimens (Dillenberger et al., 2018). Given the high number of chloroplasts in photosynthetic plant cells, chloroplast DNA molecules are in proportion much more present than nuclear DNA molecules and are therefore easier to target. For all these reasons, plastid sequences were considered as interesting markers for phylogenetic studies of taxa relationships. However, the recent advent of NGS techniques provide an unprecedented access to genome polymorphisms. Tremendous amounts of genomic data are generated for a wide range of applications, often unrelated to phylogenetics. Therefore, many taxa have now genomic data available on public gene banks which can serve for diversity analysis and therefore phylogenetics. Within few years, we switched "from famine to feast" in terms of phylogenetic markers (Hughes et al., 2006). There is currently a full scope to develop new sets of phylogenetic markers dedicated to resolve specific regions of the Tree of Life.

Phylogenomics has arisen over the last decades, especially thanks to the development of sequencing projects (Bleidorn, 2017). Phylogenomics is a portmanteau word bridging Phylogenetics and Genomics together and was invented in the late 1990s by Dr Jonathan Eisen (Eisen et al., 1997; Eisen, 1998; Eisen and Hanawalt, 1999). In its original definition, phylogenomics is used to predict gene function based on gene proximity in a context of phylogenetic inference (Eisen, 1998). Nowadays, it mostly corresponds to the use of large arrays of genome-scale sequences to resolve phylogenetic relationships between taxa (Bleidorn, 2017).

While molecular sequences for phylogenetics were so far mainly obtained through Sanger sequencing, the rapid development of NGS technologies expand the possibilities to recover large sets of molecular sequences. There now exist plenty high-throughput genomic sequencing methods to obtain thousands of sequences for phylogenomics, including microfluidic PCRs (Figure 21), Reduced representation libraries, RAD-seq, Transcriptome sequencing, Whole Genome Sequencing, Hybrid enrichment (Lemmon and Lemmon, 2013). The choice of a particular method depends on a combination of several critical factors such as (1) efficiency in non-model species, (2) flexibility in the type, size and number of target regions, (3) phylogenetic informativeness content, (4) fraction of missing data, (5) speed of data acquisition, (6) cost-effectiveness. For shallow-scale phylogenetics and phylogeography, RAD-seq was successfully applied in Bambusoideae (Wang et al., 2017a), American oaks (Hipp et al., 2014), *Carex* (Escudero et al., 2014; Massatti et al., 2016), *Pedicularis* sect. *Cyathophora* (Eaton and Ree, 2013), and *Primula tibetica* (Ren et al., 2017), improving the resolution of species complexes and intraspecific relationships despite a substantial fraction of missing data. For deeper taxonomic ranks, transcriptome sequencing recently provided unprecedented insights into the evolution of angiosperms (Leebens-Mack et al., 2019). Analysis of amplicon sequencing data provided a robust nuclear phylogeny for *Cucurbita* (Kates et al., 2017) which phylogenetic relationships were thus far appraised through the lens of plastid markers (Sanjur et al., 2002; Zheng et al., 2013). Analyses of locus obtained through hybridization-based target enrichment revealed several reticulated patterns in *Fragaria* and provided an evolutionary framework for the occurrence of allopolyploids (Kamneva et al., 2017), while reticulations in *Fragaria* phylogenies were thus far barely considered due to marker shortcomings (Potter et al., 2000; Njuguna et al., 2013; DiMeglio et al., 2014).

Regarding the analyses of phylogenomic datasets, two methods are usually developed: supermatrix and supertree (Delsuc et al., 2005; Philippe et al., 2011). The former consists in concatenating all gene sequences into a larger alignment matrix which is then analyzed using classic phylogenetic methods. The latter consists in inferring a gene tree for each gene sequence using classic phylogenetic methods and then reconcile all gene trees together in a species-tree using coalescent methods. Resorting to either of the two methods is much debated (Bininda-Emonds, 2004; Gatesy et al., 2004; Von Haeseler, 2012) but largely depends on the aim of the study and the inherent characteristics of the taxonomic group. Indeed, supermatrix tends to conceal conflicting phylogenetic signal thus resulting in highly supported topologies that might show incorrect evolutionary history (Nishihara et al., 2007; Kumar et al., 2012). On the contrary, supertrees approaches provide many perspectives on gene tree conflicts but it is then difficult to perceive the conflict origin (Galtier and Daubin, 2008). Incongruences between gene trees may actually be due to (1) violation of the orthology caused by biological factors such as incomplete lineage sorting (Figure 20), hidden paralogy or horizontal gene transfer (Maddison and Wiens, 1997), (2) stochastic error related to the relative smaller length of the genes (Galtier and Daubin, 2008), (3) systematic errors due to probabilistic model violations (Jeffroy et al., 2006).

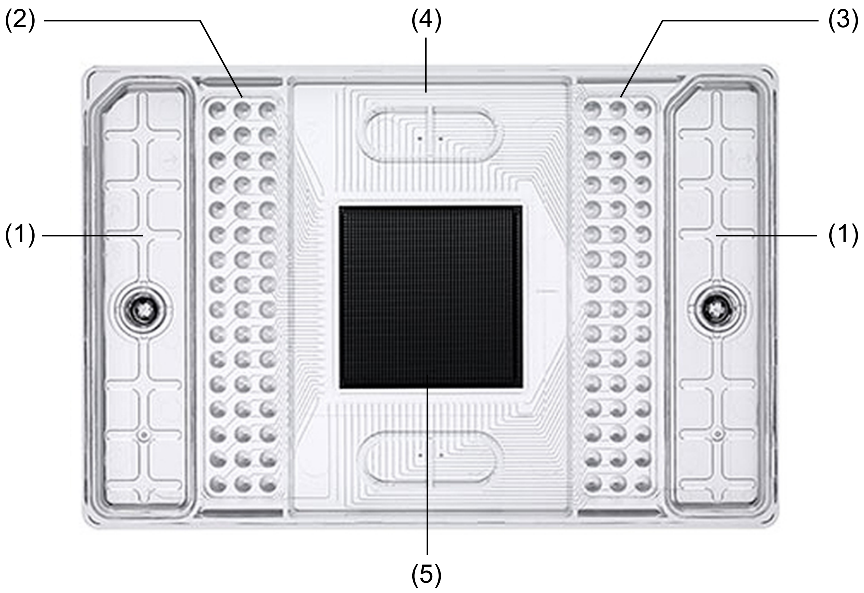


Figure 21: The 48.48 Access Array commercialized by Fluidigm. The plate allows the simultaneous amplification of 48 loci across 48 specimens, resulting in 2304 individual microfluidic PCRs. (1) compartments for PCR buffers and reagents are connected to (2) 48 wells for DNA samples and (3) 48 wells for primer pairs thanks to (4) micro canals that lead to (5) 2304 micro wells where each PCR is run individually. All the resulting amplicons are pooled and then sequenced using Illumina technologies. Barcodes are used to individualize samples.

1.4.4 Dealing with hybrids and polyploids

Assessing orthology

Using phylogenomics to resolve complex taxonomic groups still require to first identify informative nuclear loci that would be further sequenced. In this issue, much efforts are done to develop sets of low/single copy nuclear genes (SCG) able to resolve phylogenetic relationships at different levels of the Tree of Life (Li et al., 2008; Cabrera et al., 2009; Duarte et al., 2010; Liston, 2014; Deng et al., 2015). Due to their uniqueness, most SCGs that share a 1-to-1 homology with sequences present in other organisms are considered as orthologs. This mean that shared SCGs between a set of taxa are supposed to derive from a speciation event, unlike paralogous multi copy genes which may derive from a duplication event and only convey gene histories (Fitch, 1970) (Figure 22). Therefore, identifying shared 1-to-1 SCGs in a dense taxon sampling provides confidence in using othologous sequences able to trace the evolutionary history of species (Small et al., 2004).

Allele recovery

Once a set of orthologous sequences has been established within the considered taxonomic group, allele sequences must be recovered to study the parental origins and inheritance probabilities of putative subgenomes. The presence within one organism of more than one version of a SCG is a clear evidence for heterozygosity or polyploidy/genome duplications, depending on the number

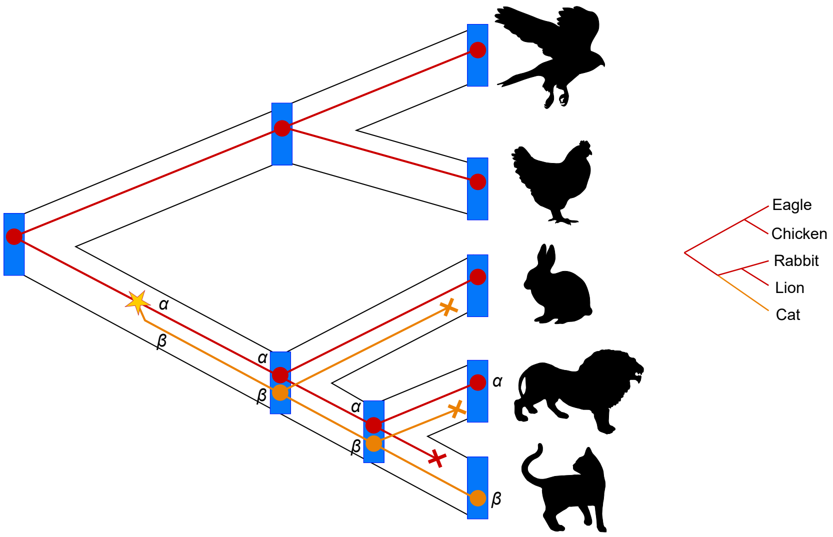
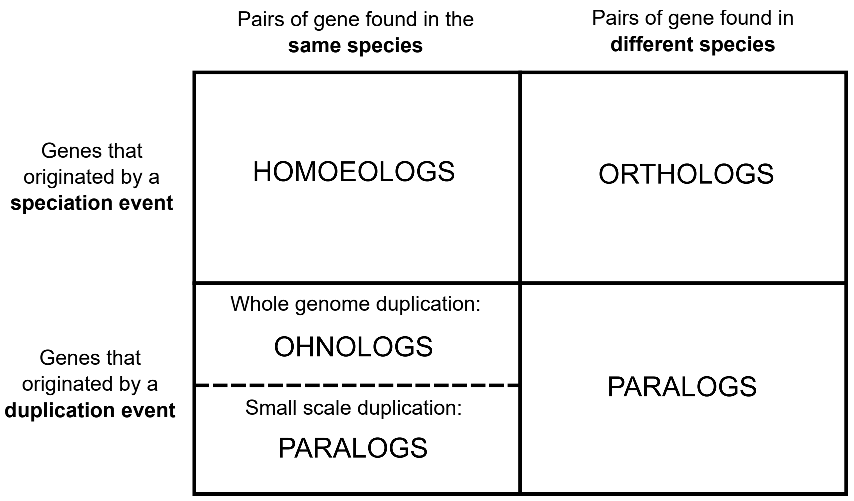


Figure 22: A schematic representation of orthology and paralogy. The black frame corresponds to the true species tree. Each blue rectangle is a speciation. The evolutionary history of a gene is represented inside the species tree topology. The gene history involves one duplication event (star) that led to two copies (α and β). α copies are orthologs (derived from a speciation event) while the β copies are paralogs (derived from a duplication event). In the course of evolution, copies were lost (crosses) in some lineages. At present, each species has only one copy of the gene thus we would consider the gene as a single-copy gene. However, due to gene loss, all the apparent single copies do not originate from a speciation event. This is called hidden paralogy. For cat, it bears the β copy while the other species have α. Therefore, the gene tree incorrectly shows that the lion is closer to the rabbit than to the cat. Adapted from slides from Fabio Pardi.

of copies that are found. In the case of polyploidy, different terminologies are given to orthologs whether they derived from a whole genome duplication within a species (ohnologs) or they resulted from interspecific hybridization associated with WGS (homoeologs/homeologs) (Figure 23). Not all NGS technologies can accommodate the recovery of allele sequences. Indeed, for WGS, DNA molecules are first sheared and the resulting fragments are sequenced using a definite read length. Many reads can contribute to the sequencing of one locus thus the resulting allele sequences may mix variations from each parental subgenome. High levels of polymorphisms inheritent to crops that extensively hybridize hamper De Bruijn Graph-based de novo assembly algorithms. This means that within one locus sequence, variations might no be phased thus hindering the possibility to study the origin of subgenomes. In transcriptome sequencing and assembly, the same issue is raised. In such heterozygous taxa, targeting locus with a definite length and that can be sequenced in one shot, either using paired-end or long read sequencing, as it is the case with amplicon sequencing seems to be the most appropriate way to recover phased variations at each locus. However, phasing allele sequences along the chromosomes is not possible with such techniques. Despite the cumbersome procedure to recover allele sequences, this information is essential for inferring a comprehensive evolutionary history since it opens the way for studying parental contribution and therefore reticulated evolutions.



	Pairs of gene found in the **same species**	Pairs of gene found in **different species**
Genes that originated by a **speciation event**	HOMOEOLOGS	ORTHOLOGS
Genes that originated by a **duplication event**	Whole genome duplication: OHNOLOGS Small scale duplication: PARALOGS	PARALOGS

Figure 23: Vocabulary associated with homologous genes. Adapted from Glover et al. (2016).

Considering reticulations

There are technical and conceptual limitations to studying reticulate evolutions on large datasets involving many genes across many species (Kamneva et al., 2017). Indeed, while generating allele sequence data for many loci becomes each year more doable and affordable as NGS methods enhance, it remains challenging and still expensive for taxonomic group with scarce available genomic resources. When allele sequence data can effectively be assembled, scientists then struggle with the analysis of allelic data for hybrid detection and analysis. This is mainly due to the fact that hybridization studies are still in their infancy and current evolutionary models rely largely on phylogenetic models that were initially developed for simple cases where only one sequence represents each species. Theories on probabilistic hybridization detection and network reconstructions actually exists and have been experimented on elementary scenarios either involving few genes or few taxa (Jin et al., 2006; Meng and Kubatko, 2009; Kubatko et al., 2009). However, these models are intractable for analyzing much larger datasets. In the case of phylogenomics on large taxonomic groups, there is currently no statistical tool of sufficient performance to produce a comprehensive network capable of clearly showing supported reticulate relationships between species (Figure 24). Therefore, choices must be taken to simplify the experimental design and reduce the complexity of the analysis. Phylogenomics is thus far restricted to simple case studies involving either diploid taxa (Arbizu et al., 2014; Kates et al., 2017) or few polyploid species (Kamneva et al., 2017). For instance, in the simple case of diploid taxonomic groups, only one consensus sequence is often considered as the representative sequence for each locus and heterozygosity is appraised through the incorporation of IUPAC ambiguity codes in the consensus sequence (Arbizu et al., 2014; Sarver et al., 2017; Kates et al., 2017). Dealing with IUPAC dismisses a clear study of parental contributions through complete and phased allele sequences, yet important to characterize hybridization patterns. Another way to reduce the complexity of studying large arrays of allele data across polyploid is to forget about phylogenomics and focus only on one or few allelic sequences to construct a multi-labeled tree (Marcussen et al., 2012, 2015; Bertrand et al., 2015). A multi-labeled tree is a tree that has several leaves labeled with the same species, each of these copies being one allele (Figure 24A). Although transforming a multi-labeled tree into a network is feasible (Huber and Moulton, 2006), there is currently no method to accommodate the combination of multiple muti-labeled trees in the presence of missing data.

All the above mentioned strategies actually enable to considerably reduce the computational burden associated with the identification of hybridizations and the calculation of their associated parameters. However, reducing the complexity is often done at the cost of accuracy. Indeed, the evolutionary history is therefore reduced to (1) a simple and smooth bifurcating tree in the case of consensus sequences or (2) a network that only detailed the evolutionary history of one or few genes with a limited extrapolation to the global evolution of species. The lack of a general phylogenomics method for considering reticulated patterns on more complex samplings leads to the development of hybrid strategies that combine classical phylogeny with small scale network inferences (Kamneva et al., 2017) and cross-checking between plastid and nuclear phylogenies (Uribe-Convers et al., 2016). These kind of strategies are currently the best way to achieve a comprehensive evolutionary history for hybrid and polyploid complexes.

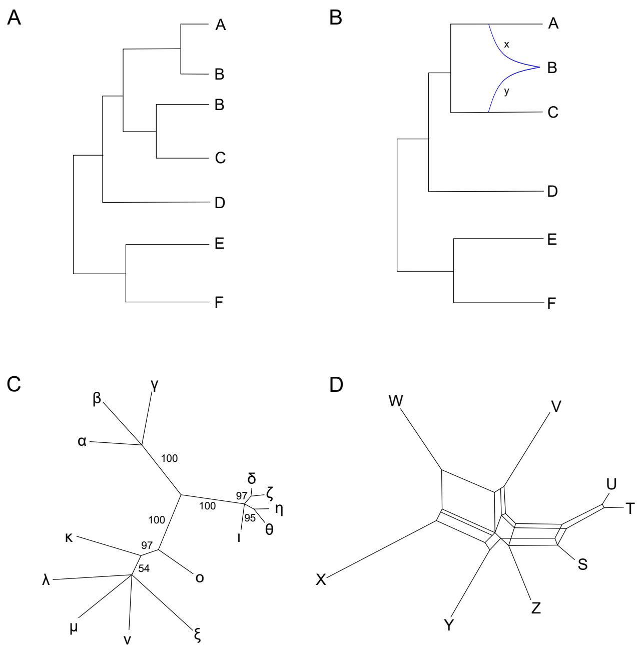


Figure 24: Examples of trees and networks. **A**. A multi-labeled tree (MUL-tree) that contains two leaf labeled with the same species (B). The MUL-tree can be converted to a hybrid network (Subfigure **B**) that highlights the reticulation leading to species B with the respective contribution of each parent (x,y). **C**. An unrooted phylogenetic tree with branch supports. **D**. A split network showing reticulated patterns.

1.5 Conclusion and thesis objectives

The complexity of wild roses has been exemplified at many differents scales, from their evolutionary history to their nomenclature, through their mating systems, and is likely the results of adaptive strategies. Thanks to interspecific hybridizations sometimes coupled with polyploidizations, wild roses succeeded in conquering most temperate and sub-tropical regions of the northern hemisphere, resulting in a large genus encompassing 100-200 species. The natural complexity of the genus has then been extended through the interest that people developed in roses. First used for their fine fragrances, roses were gradually selected for a wide range of purposes and symbols. There are currently so many cultivars and interbred varieties that it is sometimes difficult to assess the wild origin of a specimen. There will always be a grey area on the precise identification of species within *Rosa*. The evolutionary history of this genus may be better understood if it were studied at the level of sub-genera and sections. By raising the study level in this way, it seems easier to study groups of species-like individuals than to try to fit wild roses in well-delimited species concepts that do not reflect their rich evolution. Thus, the evolutionary history of wild roses will be better represented not only by a bifurcating species tree but also with networks capable of translating the reticulated relationships between the different sections and subgenera. Relationships inside each sections/subgenera could further be investigated using concepts borrowed from population genetics.

The main objective of the present thesis is therefore to reconstruct a robust and comprehensive evolutionary history of *Rosa* that is able to highlight the implication of hybridization and polyploidy in this complex genus. The present work should serve as a case study to appraise the evolutionary history of large and complex taxonomic groups. The overall aim of the thesis can thus be declined in general research questions, related to the methods:

(1) How to develop molecular markers able to meet the challenges of phylogenomics reconstructions in large groups of hybrid and polyploid taxa?

(2) What approaches can be considered to account for reticulate phylogenetic relationships within large and complex taxonomic groups?

And more specific research questions, related to the genus *Rosa*:

(3) How are the sections and subgenera of the genus *Rosa* phylogenetically related to each other?

(4) To what extent have hybridization and polyploidy shaped the genus *Rosa*?

2 Identification and assessment of variable single-copy orthologous (SCO) nuclear loci for low-level phylogenomics: A case study in the genus *Rosa*

2.1 Preamble

The first step towards phylogenomics is the identification of a substantial number of reliable phylogenetic markers. At the beginning of the thesis, there was still no reference genome sequence for the genus *Rosa*. However, an Illumina whole genome shotgun sequencing had been performed and ten thousands of scaffolds had been generated and were accessible for the heterozygous line of *Rosa* 'Old Blush'. The first idea was to use sequences already developed on Rosaceae to perform the phylogenomic analysis of the genus *Rosa*. However, by the end of the first year, half of the planned wild rose samples could not have been collected due to the withdrawal of one of the partners of the initial project. It was therefore not possible to start the DNA wet lab experiments and a massive resampling had to be scheduled. At the same time, the reference genome sequence of the haploid of *R.* 'Old Blush' was made available by the consortium, as well as dozens of unassembled wild rose genomes. I therefore decided to focus on the creation of phylogenomic markers specifically dedicated to the genus *Rosa* by using these newly obtained genomic resources. I wanted the markers to be capable of overcoming the challenges specific to the phylogeny of this genus. This allowed me to progress with my research work while letting me time to resample new specimens. This is basically how this first research chapter was developed. It is intended to be a general method, which could be applied to similar genera with available genomic resources. This work was published in BMC Evolutionary Biology and is presented in the first part of this chapter. The authors who contributed to this work are:

Kevin Debray[1], Jordan Marie-Magdelaine[1], Tom Ruttink[2], Jérémy Clotault[1], Fabrice Foucher[1] and Valéry Malécot[1].

[1] IRHS, Agrocampus-Ouest, INRA, UNIV Angers, SFR 4207 QuaSaV, Beaucouzé, France

[2] ILVO, Flanders Research Institute for Agriculture, Fisheries and Food, Plant Sciences Unit, Melle, Belgium

The contribution of each author is detailed at the end of the first part of this chapter.

In the second part of this chapter, I wanted to highlight supplementary results obtained during the thesis in relation to this chapter but which have not been published. Using the phylogenomic markers created in the first part of this chapter and resequencing data, I wanted to clarify the origins of *Rosa chinensis* 'Old Blush'.

2.2 Introduction

Next-Generation Sequencing (NGS) methods are now extensively used to address various scientific issues ranging from ecology to medicine, and become more affordable each year. Molecular phylogenetic studies greatly benefit from the high-throughput sequencing technologies that generate a wealth of information to decipher taxa relationships (Straub et al., 2012). The 1000 plant (1KP) project (Matasci et al., 2014) released large-scale gene sequencing data for over 1000 species, and thousands of other genome sequences are expected in the near future (Cheng et al., 2018).

Relationships among angiosperms are relatively well-known, ranging from deep branches to the family rank (Soltis et al., 2011; Stevens, 2017), with some exceptions (Refulio-Rodriguez and Olmstead, 2014). However, it is often challenging to understand shallower relationships in particular angiosperm families, especially between species (Hughes et al., 2006; Lyu et al., 2018). Rapid diversifications are common to angiosperms, involving evolutionary processes such as polyploidization and hybridization (Soltis and Soltis, 2009; Ren et al., 2018). These two processes are likely to occur between closely-related species, generally inside genera (Mallet, 2005). While plant molecular phylogenetics has long been dominated by plastid sequence analysis (Shaw et al., 2005; Gitzendanner et al., 2018), identifying nuclear genes has now become an important issue in phylogenetic reconstruction, especially for hybrid and polyploid taxa (Babineau et al., 2013). Nuclear markers generally show higher rates of evolution than plastid sequences and may contain more informative nucleotide substitutions to distinguish closely-related taxa (Sang, 2002). Whereas plastid genomes are mainly maternally inherited in angiosperms (Reboud and Zeyl, 1994), nuclear markers contain sequence signatures of both parents, making them more useful to study hybridization and polyploidization events in taxa at the boundary between species and populations (Sang, 2002; Joly et al., 2006b). Up to now, only few nuclear genes that are ubiquitously present in species across the Tree of Life have been commonly used for phylogenetics such as nuclear ribosomal internal transcribed spacers (nrITS) and glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*). However, such sequences may present multiple issues for phylogenetic analyses. *GAPDH* is better suited to resolve relationships at the kingdom or class level (Canback et al., 2002; Martin and Cerff, 2017) than at the genus or species levels. nrITS exist in multiple copies that might not evolve at the same rate so that comparison between them may mislead phylogenetic analyses (Poczai and Hyvönen, 2010; Naumann et al., 2011).

With the ever-growing number of available whole genome sequences, several sets of new nuclear markers have been published to help unravel phylogenetic relationships at different plant taxonomic levels, ranging from the angiosperm clade (Li et al., 2008; Duarte et al., 2010; Han et al., 2014; Liu et al., 2017) to particular families (Cabrera et al., 2009; Liston, 2014; Lemmon and Lemmon, 2012). Specific attention has been given to single-copy genes (SCG) that go beyond the issues of conventional markers (ie plastid sequences or ubiquitous nuclear genes) and turn out to be good candidates for phylogenetic analysis (Sang, 2002; Small et al., 2004). In addition to their biparental inheritance and their high content of informative characters, SCGs ease the identification of orthologs (Sang, 2002). Orthologs are genes that derive from speciation events, as opposed to paralogs that derive from duplication events and should therefore be discarded from phylogenetic analyses. Consequently, sequences found in a wide range of taxa and that share a 1-to-1 homology with core SCGs may have resulted from speciation events and may therefore be considered as orthologous sequences. In angiosperm genomes, 8-35% of the genes are found as a single copy (Han et al., 2014), providing the opportunity to find many orthologous sequences well suited to carrying out phylogenetic studies at various taxonomic levels.

Phylogenomics, i.e., the use of large arrays of genome sequences to infer phylogenetic relationships, has emerged over the last few years and is increasingly used in molecular studies of taxa relationships (Roure et al., 2007; Bleidorn, 2017). With the tremendous increase in plant genome sequencing projects (Anonymous, 2018), it is now feasible to include thousands of sequences for phylogenetic analysis. Since a larger set of genomic sequences are included in the comparison, topological conflicts between individual gene trees and the species-tree arise (Maddison and Wiens, 1997; Jeffroy et al., 2006; Prasad et al., 2008; Degnan and Rosenberg, 2009). These conflicts could be due to horizontal gene transfer, incomplete lineage sorting, and gene duplication and gene loss (Patané et al., 2018). To circumvent these particular issues, a common method consists in concatenating the gene sequences, assuming that the true overall phylogenetic signal would arise and conceal the noise contained in individual genes (Bapteste et al., 2005; Von Haeseler, 2012). Several methods have been developed to assess this noise and to help in selecting the best marker set, with the most informative characters captured with the lowest number of sequences. Most of these methods rely on distance metrics derived from tree topologies (Robinson and Foulds, 1981) and branch length comparisons (Kuhner and Felsenstein, 1994; Farris et al., 1995) or, alternatively, on likelihood ratio tests (Huelsenbeck and Bull, 1996; Waddell et al., 2000) combined with various clustering methods (Planet and Sarkar, 2005; Leigh et al., 2008, 2011; Gori et al., 2016; Narechania et al., 2016). Other methods use a conceptual index to assess the phylogenetic utility of sequences (Townsend, 2007). The main goal of marker selection is to find the optimal balance between character sampling and taxon sampling. Too few markers may lead to inaccurate estimations of phylogenetic relationships whereas too many markers increase the computational needs and the overall cost of the experiment, especially for phylogenomic studies involving a broad number of taxa.

Phylogenetic analysis of the genus *Rosa* is challenging because the genus comprises approximatively 150 species distributed in the Northern Hemisphere that are the result of a complex evolutionary history involving multiple hybridization and polyploidization events across the last 30 M years (Fougère-Danezan et al., 2015). Currently, Rehder's classification (Rehder, 1940), slightly modified by Wissemann (2003a), is still used and divides the genus into four subgenera (*R.* subgen. *Rosa*, *R. subgen. Hulthemia* (Dumort.) Focke, *R.* subgen. *Platyrhodon* (Hurst) Rehder and *R.* subgen. *Hesperhodos* Cockerell). About 95% of the wild rose species belong to the subgenus *Rosa* which is further divided into ten sections (*R.* sect. *Pimpinellifoliae* (DC.) Ser., *R.* sect. *Gallicanae*, *R.* sect. *Caninae* (DC.) Ser., *R.* sect. *Carolinae* Crép., *R.* sect. *Rosa* [= *R.* sect. *Cinnamomeae* (DC.) Ser.], *R.* sect. *Synstylae* DC., *R.* sect. *Chinenses* Ser. [= *R.* sect. *Indicae* Thory], *R.* sect. *Banksianae* Lindl., *R.* sect. *Laevigatae* Thory and *R.* sect. *Bracteatae* Thory). In this paper, we adopt the designation of *Rosa cinnamomea* L. (syn. *Rosa majalis* Herrm.) as the type species of the genus, a proposal from Jarvis (1992) and validated in 2005 at the Vienna International Botanical Congress. This implies that the section previously known as *Rosa* sect. *Cinnamomeae* (DC.) Ser. is renamed *R.* sect. *Rosa*. In addition, Wissemann (2003a) subdivided the *R.* sect. *Caninae* into six subsections (*R.* subsect. *Trachyphyllae* H. Christ, *R.* subsect. *Rubrifoliae* Crép., *R.* subsect. *Vestitae* H. Christ, *R.* subsect. *Rubiginae* H. Christ., *R.* subsect. *Tomentellae* H. Christ and *R.* subsect. *Caninae*). *R.* sect. *Caninae* is an evidence of rapid radiation in the genus *Rosa*. While this section accounts for approximatively 20% of the *Rosa* species, it appeared only ca. 6 MYa (Fougère-Danezan et al., 2015). Thus far, the phylogenetic relationships among wild roses have been explored with nrITS (Iwata et al., 2000; Matsumoto et al., 2000b,a; Wu et al., 2001; Wissemann and Ritz, 2005; Qiu et al., 2012), chloroplast regions (Matsumoto et al., 1998; Wissemann and Ritz, 2005; Bruneau et al., 2007; Qiu et al., 2012, 2013; Kellner et al., 2014; Liu et al., 2015; Fougère-Danezan et al., 2015), and *GAPDH* (Joly et al., 2006b; Fougère-Danezan et al., 2015), as phylogenetic markers. The phylogenetic relationships derived from these conventional markers either focused on specific sections, or were poorly resolved, because many clades lacked support due to little sequence variation between the sampled species. Nevertheless, Fougère-Danezan et al. (2015) distinguished three main clades (sect. *Synstylae* and allies, sect. *Pimpinellifoliae*, and sect. *Cinnamomeae* [i.e., sect. *Rosa*] and allies) and is currently the most complete and resolved phylogeny of the genus *Rosa*. The recent publication of a high-quality reference genome sequence of *Rosa* 'Old Blush' (Raymond et al., 2018; Hibrand Saint-Oyant et al., 2018), a putative hybrid between *R. chinensis* and *R. odorata* var. *gigantea* (Meng et al., 2011), provides an excellent resource to mine for nuclear sequences for high-resolution phylogenomic analysis of the genus *Rosa*. Moreover, multiple poor quality draft genomes of wild *Rosa* species have recently been released and can also be mined for shared loci with sequence variations between the different species (Table 2). We used these genomes here to present a general method to identify a set of single-copy nuclear orthologous loci that can be amplified from species across the genus. These sequences contain the sequence variations required to study species relationships through phylogenomics. The method was developed for the genus *Rosa*, and can be used at different taxonomic levels and groups.

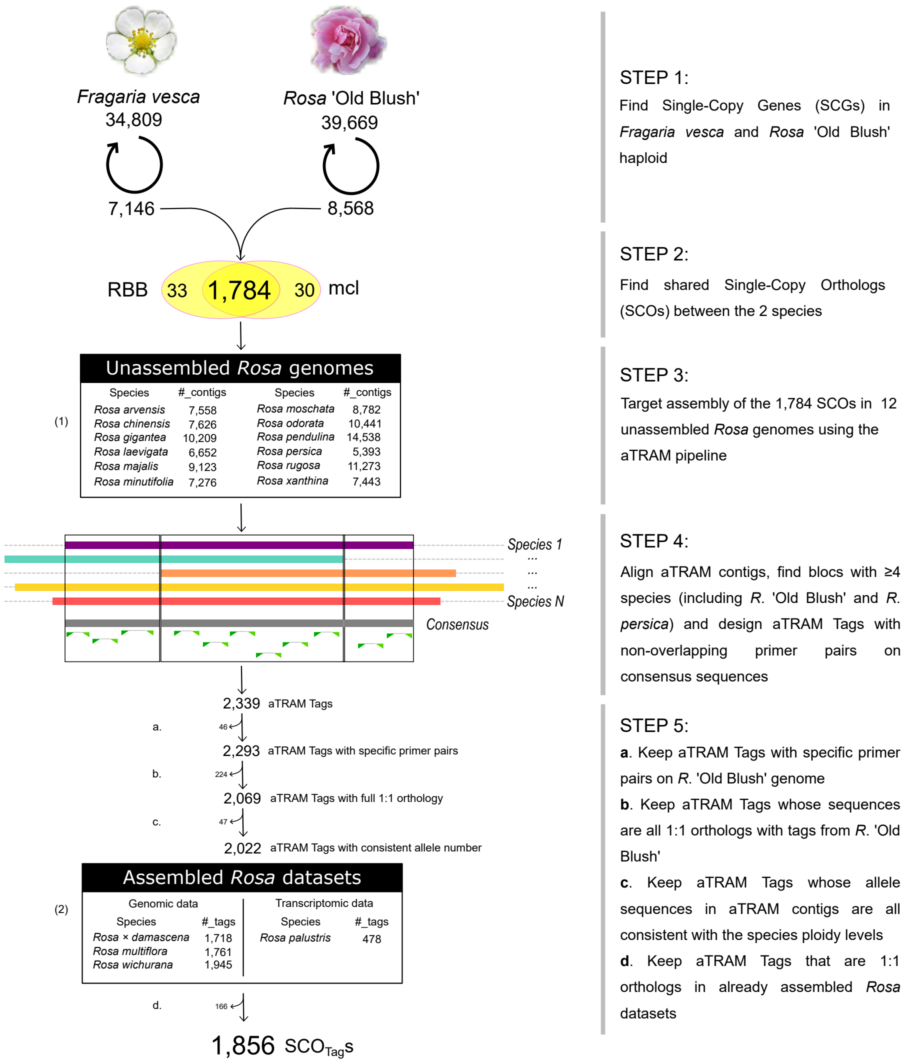
2.3 Material and methods

2.3.1 Identification of single-copy orthologs in *Rosa* 'Old Blush' and *Fragaria vesca*

Single-copy nuclear genes were identified by comparing annotated protein sets in the haploid reference genome sequences of *Rosa* 'Old Blush' (Hibrand Saint-Oyant et al., 2018) and *Fragaria vesca* (Shulaev et al., 2011). First, the annotated protein set from each genome was compared to itself using an all-against-all BLAST+ (Camacho et al., 2009) search. Outputs were parsed using the tcl script (Kozik et al., 2005) with an e-value cutoff of $1e^{-}10$, identity of at least 30% and coverage above 70% of the query. Single-copy nuclear genes were identified as those with a unique blast hit to themselves (Step 1, Figure 25). Next, two methods were used to identify single-copy orthologs (SCOs) shared between *Rosa* 'Old Blush' and *F. vesca*. In the first method, a reciprocal best-hit blast (RBB) was performed between *Rosa* 'Old Blush' and *F. vesca* sets of single-copy genes, and SCOs were identified as pairs of proteins with each other as the best scoring match in the respective genome. Second, the Markov clustering algorithm (mcl) method (van Dongen, 2012) was run via the mclblastline command (Enright et al., 2002) to cluster all single-copy proteins from *Rosa* 'Old Blush' and *F. vesca* into groups using an inflation value of intermediate stringency (3.0). Genes found as SCOs in both methods were retained for downstream analysis (Step 2, Figure 25). A synteny analysis was also performed to compare the position of the SCOs in the genome assemblies of *F. vesca* (Edger et al., 2018) and *R.* 'Old Blush', and to further assess the orthology assumption. Finally, we also used BLAST with the above settings to compare our set of SCOs to three published ortholog sets to evaluate the redundancy of our SCOs (957 *Arabidopsis*-*Populus*-*Vitis*-*Oryza* (APVO) single-copy genes (Duarte et al., 2010), 257 Low-Copy Nuclear Genes for Rosaceae phylogenomics (LCNG) (Liston, 2014) and 1041 Rosaceae Conserved Ortholog Set of markers (RosCOS) (Cabrera et al., 2009)).

2.3.2 Reconstruction of nuclear SCOs and plastid loci in *Rosa* sp.

To identify sequence variations within the SCOs across the genus *Rosa*, we retrieved the corresponding sequences from already published whole genome shotgun (WGS) Illumina paired-end sequence data of 16 *Rosa* species (Table 2 and Figure 25). For 12 unassembled genomes, WGS reads were processed with the aTRAM v1.0 iterative pipeline (Allen et al., 2015) to assemble the SCOs. Briefly, reads are first assigned to partitions, also called shards, to ease the pipeline parallelization and to optimize the computing needs. Second, a SCO protein sequence is used as a query to retrieve homologous reads through a BLASTX search against shards. Corresponding forward or reverse reads are then retrieved for ABySS v2.0 assembly (Jackman et al., 2017). Assembled contigs are iteratively used as queries for the next round of assembly. As a result, contig length increases and this iterative process may lead to the assembly of the entire SCO locus, including introns and untranslated regions. We performed three iterations of assembly on


Fragaria vesca
34,809
7,146
Rosa 'Old Blush'
39,669
8,568
RBB
33
1,784
30
mcl
Species 1
...
...
Species N
Consensus
2,339 aTRAM Tags
a.
46
2,293 aTRAM Tags with specific primer pairs
b.
224
2,069 aTRAM Tags with full 1:1 orthology
c.
47
2,022 aTRAM Tags with consistent allele number
d.
166
1,856 $\mathrm{SCO_{Tag}s}$

(1) Unassembled *Rosa* genomes

Species	#_contigs	Species	#_contigs
Rosa arvensis	7,558	*Rosa moschata*	8,782
Rosa chinensis	7,626	*Rosa odorata*	10,441
Rosa gigantea	10,209	*Rosa pendulina*	14,538
Rosa laevigata	6,652	*Rosa persica*	5,393
Rosa majalis	9,123	*Rosa rugosa*	11,273
Rosa minutifolia	7,276	*Rosa xanthina*	7,443

(2) Assembled *Rosa* datasets

Genomic data		Transcriptomic data	
Species	#_tags	Species	#_tags
Rosa × damascena	1,718	*Rosa palustris*	478
Rosa multiflora	1,761		
Rosa wichurana	1,945		

STEP 1:
Find Single-Copy Genes (SCGs) in *Fragaria vesca* and *Rosa* 'Old Blush' haploid

STEP 2:
Find shared Single-Copy Orthologs (SCOs) between the 2 species

STEP 3:
Target assembly of the 1,784 SCOs in 12 unassembled *Rosa* genomes using the aTRAM pipeline

STEP 4:
Align aTRAM contigs, find blocs with ≥4 species (including *R.* 'Old Blush' and *R. persica*) and design aTRAM Tags with non-overlapping primer pairs on consensus sequences

STEP 5:
a. Keep aTRAM Tags with specific primer pairs on *R.* 'Old Blush' genome
b. Keep aTRAM Tags whose sequences are all 1:1 orthologs with tags from *R.* 'Old Blush'
c. Keep aTRAM Tags whose allele sequences in aTRAM contigs are all consistent with the species ploidy levels
d. Keep aTRAM Tags that are 1:1 orthologs in already assembled *Rosa* datasets

Figure 25: Data-mining worflow to identify single-copy orthologous tags ($\mathrm{SCO_{Tag}s}$) for phylogenomics. Single-copy genes (SCGs) from reference genomes are identified using a self-blast procedure (step 1). The two SCG sets are compared to each other to retrieve shared single-copy orthologs (SCOs) (step 2). SCOs are target-assembled from unassembled whole genome shotgun sequencing data using the aTRAM pipeline. Numbers presented in table (1) correspond to the total number of contigs that were assembled for each *Rosa* species with an unassembled genome (step 3). Contig sequences from each SCO are aligned using mafft and the resulting alignment is sliced in regions $\geq$ 300 bp covered by $\geq$ 4 taxa including *Rosa* 'Old Blush' and *Rosa persica*. For each region, pairs of primers are designed on the consensus sequence and the most variable non-overlapping $\mathrm{SCO_{Tag}s}$ are retained (step 4). Additional filtering steps enables to discard $\mathrm{SCO_{Tag}s}$ with unspecific primer pairs (step 5a), $\mathrm{SCO_{Tag}s}$ that do not pass the RBB test of orthology (5b), $\mathrm{SCO_{Tag}s}$ with inconsistent number of alleles regarding the genome ploidy level (5c) and to find $\mathrm{SCO_{Tag}s}$ in whole genome shot gun assemblies of three additional *Rosa* species (step 5d) and seven outgroups. Numbers in table (2) correspond to the number of $\mathrm{SCO_{Tag}s}$ that were retrieved for each of the four *Rosa* species with already assembled datasets. The procedure is described in detail in the Methods section. RBB: Reciprocal Best Blast; mcl: Markov CLuster algorithm.

the GenoToul bioinformatics high-performance computing cluster using 16 cores of Intel® Xeon® computers with a 2.50GHz processor. For each SCO in each unassembled *Rosa* genome, the contig with the highest alignment score on the *Rosa* 'Old Blush' reference SCO sequence was selected as the representative orthologous sequence for this genome (Step 3, Figure 25). Then, for each SCO, we created mafft (Katoh and Standley, 2013) alignments between all orthologous sequences. Alignments were screened to find regions covered by at least four taxa, including *Rosa* 'Old Blush' and *Rosa persica*, considered as the most divergent *Rosa* taxon (Fougère-Danezan et al., 2015) and even considered to be in a separate genus in former classifications (Dumortier, 1824; Robyns, 1938). Strict consensus sequences of these regions were used to design generic conserved primer pairs with Primer3 (Untergasser et al., 2012), so that any fragment could further be amplified using the polymerase chain reaction technique (Step 4, Figure 25). Conditions for primer design were a melting temperature between 59°C and 61°C, a maximal homo-polymer of 3, at least one 3'-GC clamp and amplicon size between 300 bp and 550 bp. For each SCO, a maximum of 100 primer pairs spanning the entire consensus sequence were designed. We then selected the most variable non-overlapping amplicon tags for each region and checked for specificity of their corresponding primers on the haploid reference genome of *Rosa* 'Old Blush' (Step 5a, Figure 25). We additionally retrieved positional information on untranslated transcribed region (UTR), intron, and exon locations for each tag. To further assess the orthology assumption of the tags, we ran additional tests. First, we checked that each tag has a number of alleles in assembled aTRAM contigs that is compatible with the species genome ploidy level (Step 5b, Figure 25). Then, we subjected each targeted tag sequence to a reciprocal-best BLAST (Step 5c, Figure 25). Since the sequence of *R.* 'Old Blush' was used as the query for the aTRAM assembly, blasting each targeted tag sequence back to the genome of *R.* 'Old Blush' provide the RBB test for orthology. Any tag sequence that did not pass these two tests led to the rejection of all *Rosa* sequences associated with this tag for downstream analysis. Finally, the corresponding tag were retrieved from recently assembled genomes of three additional *Rosa* species (Table 2) (Step 5d, Figure 25). We also used an assembled transcriptome of *Rosa palustris* (Johnson et al., 2012) because it belongs to the *Rosa* sect. *Carolinae* and is related to several wild roses native to North America. Only exonic tag can be retrieved from transcriptome sequencing data of *R. palustris*. We used a BLAST search to retrieve tag sequences from assembled genomes/transcriptome of *Rosa* species (e-value $\leq$ 1e-10; identity $\geq$ 65%; 100% coverage of the consensus query tag; maximum query-subject length difference of ±20%). If multiple best hits were found, we arbitrary choose one of them as the representing sequence for the *Rosa* species. If the number of best hits was not consistent with the ploidy level of the *Rosa* species genome, we discarded all sequences related to this tag for downstream analysis. We additionally checked that edges of retrieved sequences corresponded to primer pairs. Thanks to the different filtering procedures that we applied on the initial set of tags, we considered that the resulting tags are Single-Copy Orthologous tags (SCO_{Tag}s), suited to reconstructing phylogenomics relationships in the genus *Rosa*. We applied the same procedure as Step 5d, Figure 25 to identify similar SCO_{Tag} sequences in seven sister outgroups belonging to the subfamily Rosoideae (*Rubus occidentalis*, *Fragaria vesca*, *Fragaria iinumae*, *Fragaria nipponica*, *Fragaria nubicola*, *Geum urbanum*, and *Potentilla micrantha*) (Table 2), except that we did not check that edges of sequences strictly corresponded to the respective primer pairs and that we did not discard all SCO$_{\text{Tag}}$ sequences if the number of best hits was not consistent with the ploidy level of the outgroup species genome.

The same procedure was applied to retrieve three plastid sequences (psbA-trnH, trnG and trnL) from (un)assembled genome sequences of the same *Rosa* species. When the procedure failed to assemble plastid sequences, we retrieved corresponding plastid sequences from NCBI GenBank (Appendix A, Supplementary table A.1).

2.3.3 Assessment of phylogenomic utility

SCO$_{\text{Tag}}$ sequences from different species were aligned using mafft (Katoh and Standley, 2013), and Gblock (Castresana, 2000) was used to trim poorly aligned regions. Variable and parsimony-informative site (PIS) contents were calculated per SCO$_{\text{Tag}}$ alignment. Gaps were treated as a fifth base. We then computed phylogenetic informativeness (PI) per SCO$_{\text{Tag}}$ using the formula presented in Townsend and Leuenberger (2011) and implemented in the PhyDesign online application (López-Giráldez and Townsend, 2011). For this analysis, all SCO$_{\text{Tag}}$ alignments were concatenated into a super-matrix and the best partition scheme was searched with PartitionFinder v2.1.1 (Lanfear et al., 2014). The partitioned matrix served to construct a Maximum Likelihood (ML) species-tree using RAxML v8.1.5 (Stamatakis, 2014). The species-tree was then converted to a chronogram in R using the function chronos in the package ape (Paradis et al., 2017) by applying one calibration point on the crown node of *Rosa*, dated at 30 MYa (Fougère-Danezan et al., 2015). We uploaded the partition concatenated matrix and the chronogram on PhyDesign. Substitution rates were calculated for each SCO$_{\text{Tag}}$ in HyPhy (Pond et al., 2005) using the best generalized time-reversible (GTR) model, with empirical base frequencies, found for the super-matrix in jModeltest2 (Darriba et al., 2012). Some SCO$_{\text{Tag}}$ alignments have sites for which substitution rate was incorrectly determined leading to high spikes close to time 0. Since those high spikes have no real biological meaning and correspond to artefacts, we decided to remove them. To do so, we first identify SCO$_{\text{Tag}}$ with such spikes by looking for SCO$_{\text{Tag}}$ PI profiles with more than 1 maximum. Second, we retrieved the estimated substitution rates for each SCO$_{\text{Tag}}$ with high spikes and looked for an elbow in the distribution of substitution rates. The substitution rate found at the elbow served as a threshold to discard sites with unusual substitution rate. We repeated this second step one time to totally remove high spikes from PI profiles. The python script (PhantomSpikesRemover.py) that we developed to trim SCO$_{\text{Tag}}$ PI profiles and alignments is available at `https://github.com/kdebray/SCOtags`. The same procedure was applied to the three plastid loci to recover their PIS content and PI profiles.

To further determine the underlying phylogenetic conflicts between SCO$_{\text{Tag}}$s, we looked for well-supported incongruences between SCO$_{\text{Tag}}$ tree topologies. For each SCO$_{\text{Tag}}$ alignment with at least one outgroup sequence, we determined the best nucleotide substitution model using jModelTest v2

Table 2: References used for Whole Genome Shotgun data. (†) indicates unassembled Whole Genome Shotgun data. *IRHS* Institut de Recherche en Horticulture et Semences, *ENS* Ecole Normale Supérieure, *ILVO* Instituut voor Landbouw-, Visserij- en Voodingsonderzoek, *NCGR* National Clonal Germplasm Repository

	Species	Ploidy of the genome sequence	Sample origin	BioProject/ SRA code	Original publication
Ingroup	*Rosa* 'Old Blush'	1x	IRHS, Beaucouzé, France	-	Hibrand Saint-Oyant et al. (2018)
	†*Rosa arvensis* Huds.	2x	Jardin expérimental de Colmar, Colmar, France	SRX3286288	Raymond et al. (2018)
	†*Rosa chinensis* Jacq. var. *spontanea* (Rehd. & Wils.) T. T. Y & T. C. Ku	2x	Roseraie du Val-de-Marne, L'Hay-les-Roses, France	SRX4006790	Hibrand Saint-Oyant et al. (2018)
	Rosa × *damascena* Mill.	4x	Bulgaria	PRJNA322107	-
	†*Rosa gigantea* Collet ex Crép.	2x	Lyon botanical garden, Lyon, France	SRX3286284, SRX3286283	Raymond et al. (2018)
	†*Rosa laevigata* Michx.	2x	Roseraie du Val-de-Marne, L'Hay-les-Roses, France	SRX4006792	Hibrand Saint-Oyant et al. (2018)
	†*Rosa majalis* Herrm.	2x	ENS Lyon, Lyon, France	SRX3286287	Raymond et al. (2018)
	†*Rosa minutifolia* var. *alba* Engelm.	2x	Roseraie du Val-de-Marne, L'Hay-les-Roses, France	SRX4006787	Hibrand Saint-Oyant et al. (2018)
	†*Rosa moschata* Herrm.	2x	Roses Loubert rose garden, Les-Rosiers-sur-Loire, France	SRX4006793	Hibrand Saint-Oyant et al. (2018)
	Rosa multiflora Thunb. ex Murr.	2x	Keisei Rose Nurseries, Chiba, Japan	PRJDB4738	Nakamura et al. (2017)
	†*Rosa odorata* (Andr.) Sweet	2x	Lyon botanical garden, Lyon, France	SRX3286293	Raymond et al. (2018)
	Rosa palustris Marsh.	2x	NA	ERS1829481	Johnson et al. (2012)
	†*Rosa pendulina* L.	2x	Lyon botanical garden, Lyon, France	SRX3286278	Raymond et al. (2018)
	†*Rosa persica* Michx. ex Jussieu	2x	Roses Loubert nurseries, Les-Rosiers-sur-Loire, France	SRX4006789	Hibrand Saint-Oyant et al. (2018)
	†*Rosa rugosa* Thunb.	2x	Roseraie du Val-de-Marne, L'Hay-les-Roses, France	SRX4006791	Hibrand Saint-Oyant et al. (2018)
	Rosa wichurana Crép.	2x	ILVO, Melle, Belgium	PRJNA504542	-
	†*Rosa xanthina* var. *xanthina* f. *spontanea* Rehd.	2x	Roses Loubert rose garden, Les-Rosiers-sur-Loire, France	SRX4006788	Hibrand Saint-Oyant et al. (2018)
Outgroup	*Fragaria vesca* L.	1x	NCGR, Corvallis, OR, USA	PRJNA66853	Shulaev et al. (2011)
	Fragaria iinumae Makino	2x	Kagawa University, Kagawa, Japan	PRJDB1478	Hirakawa et al. (2014)
	Fragaria nipponica Makino	2x	Kagawa University, Kagawa, Japan	PRJDB1479	Hirakawa et al. (2014)
	Fragaria nubicola Lindl. ex Lacaita	2x	NCGR, Corvallis, OR, USA	PRJDB1480	Hirakawa et al. (2014)
	Geum urbanum L.	6x	Punnets Town, UK	PRJEB23412	Jordan et al. (2018)
	Potentilla micrantha Ramond ex DC.	2x	Avala, Serbia	PRJEB18433	Buti et al. (2018)
	Rubus occidentalis L.	2x	Rich Mountain, South Carolina, USA	-	VanBuren et al. (2016)

(Darriba et al., 2012), and we estimated corresponding ML tree with PhyML (Guindon et al., 2003). We then used PhyParts (Smith et al., 2015) to map resulting $\mathrm{SCO}_{\mathrm{Tag}}$-trees onto the species-tree topology, previously obtained by concatenation of all $\mathrm{SCO}_{\mathrm{Tag}}$s followed by a ML tree estimation. Briefly, each gene-tree is rooted on outgroup species and then split into bipartitions that are compared to all bipartitions present in the species-tree. A gene-tree bipartition h is concordant with a species-tree bipartition s if all of the ingroup of h is included in the ingroup of s and if all of the outgroup of h is included in the outgroup of s (Smith et al., 2015). We applied a bootstrap filter of 70% so that only medium to well-supported bipartitions are taken into account for the concordance calculations. As a result, each node of the species-tree is labeled with the fraction of concordant $\mathrm{SCO}_{\mathrm{Tag}}$s and conflicting $\mathrm{SCO}_{\mathrm{Tag}}$s. In addition, we used Astral (Zhang et al., 2018a) v5.6.3 with default parameters to build a coalescent species-tree from the $\mathrm{SCO}_{\mathrm{Tag}}$ trees and to compute Local Posterior Probabilities associated with each quadripartitions of the coalescent species-tree. We also calculated the Internode Certainty All (ICA) for each node of the species-tree topology, as implemented in PhyParts. ICA values near 0 indicate major conflicts with similar frequencies among conflicting bipartitions. ICA values near 1 indicate a strong certainty in the bipartition, meaning that few alternative bipartitions with low frequencies have been found. Although ICA score is not directly comparable to bootstrap support (BS), it provides more information about the distribution of conflicts among phylogenomic loci for a specific bipartition (Smith et al., 2015). In addition, we also summarized topological conflict between $\mathrm{SCO}_{\mathrm{Tag}}$ trees through a species network. For this analysis, we first collapsed branches that were poorly supported (ie. $\mathrm{BS} < 70$ %) using a custom R script and the function *di2multi* in the ape package. Then, we combined all clean $\mathrm{SCO}_{\mathrm{Tag}}$ trees in a FilteredSuperNetwork as implemented in SplitsTree (Huson and Bryant, 2006) v4.

2.4 Results

2.4.1 Identification of single-copy orthologs (SCO) in *Rosa* 'Old Blush' and *Fragaria vesca*

We compared annotated proteins from reference genomes of haploid *Rosa* 'Old Blush' (Hibrand Saint-Oyant et al., 2018) and *Fragaria vesca* (Shulaev et al., 2011) to identify single-copy orthologs (SCOs) using the all-against-all BLAST+ procedure. We found that *Rosa* 'Old Blush' (resp., *Fragaria vesca*) has 8568 single-copy genes (resp., 7146), which represents 21.6% (resp., 20.5%) of all predicted proteins for this genome (Step 1, Figure 25).

Using these two sets of single-copy genes, the Reciprocal Best Blast (RBB) procedure identified 1817 shared SCOs between *Rosa* 'Old Blush' and the Markov Clustering (mcl) identified 1814 shared SCOs. A total of 1784 SCOs were commonly identified by both methods (Step 2, Figure 25). These common SCOs are evenly distributed across the seven chromosomes of the haploid genome of *Rosa* 'Old Blush' (Figure 26A). The synteny analysis reveals that the order of SCOs along the genome of *Fragaria vesca* and *R.* 'Old Blush' is well conserved (Appendix A, Supplementary figure A.1). The great majority (73%) of SCOs that we found are new and were never published before in other ortholog sets (Appendix A, Supplementary figure A.2).

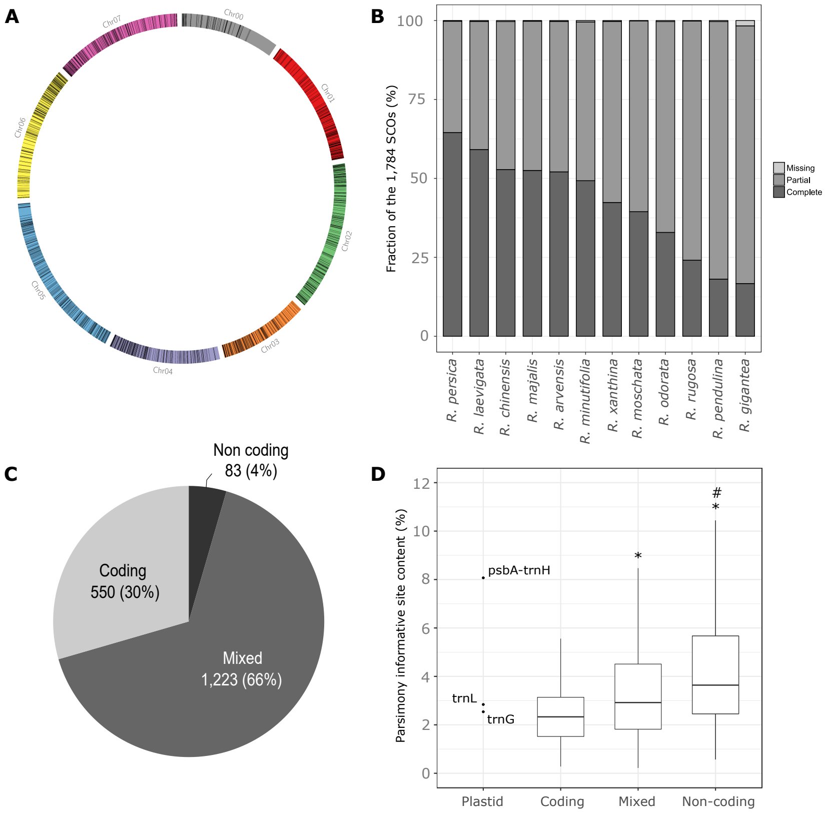


Figure 26: Characterization of the plastid loci and nuclear $\mathrm{SCO_{Tag}}$s. **A**. Position of the 1784 single-copy orthologs (SCOs) in the seven pseudo chromosomes and unanchored scaffolds (Chr00) of the haploid genome sequence of *Rosa* 'Old Blush'. **B**. Completeness of SCOs in the 12 unassembled rose genomes. *Missing* means that no contig matching the reference SCO could have been assembled; *partial* means that only part of the reference SCO was assembled; *complete* means that the complete reference SCO is covered by at least one assembled contig. **C**. Structural annotation of 1856 $\mathrm{SCO_{Tag}}$s. **D**. Parsimony-informative site (PIS) content for plastid sequences (psbA-trnH, trnL and trnG) and the nuclear $\mathrm{SCO_{Tag}}$s. $\mathrm{SCO_{Tag}}$s are divided into three categories: coding regions (exons), non-coding (untranslated regions and introns), and mixed regions (containing both coding and non-coding regions). ($*$) and ($\#$) denote significant differences between coding and mixed regions and between mixed and non-coding regions, respectively (t-test; p-value < 0.05).

2.4.2 Target assembly and primer design

We applied the automated Target Restricted Assembly Method (aTRAM) for the 1784 selected SCOs to reconstruct (either partly or completely) their corresponding orthologs from the available unassembled genome sequences of 12 *Rosa* species (Table 2). A mean of 1776 SCOs (ranging from 1754 SCOs for *R. gigantea* to 1782 SCOs for *R. moschata*) was retrieved per Rosa species (Figure

26B).

After creating alignments of the aTRAM contigs for each of the 1784 SCOs, we were able to identify 2874 sub-alignments of at least 300 bp that were covered by at least four taxa, including the haploid reference genome of *Rosa* 'Old Blush' and the most divergent species *R. persica*. Strict consensus sequences of these sub-alignments were used to design a total of 2339 in silico primer pairs flanking variable non-overlapping tags of 300-550 bp. A total of 1000 out of the 1784 SCOs have at least one tag, with an average of 2.3 tags per SCO (ranging from 1 to 14). Of the 2339 candidate tags, 483 did not pass the post-assembly tests (Step 5, Figure 25). In details, 46 tags were removed due to unspecific binding of their primer pairs to the haploid reference genome sequence of *Rosa* 'Old Blush'; 224 tags did not pass the RBB test of orthology; 47 tags did not have a consistent allele number in aTRAM contigs regarding the ploidy level of the unassembled *Rosa* genome; 166 tags did not have a consistent hit number regarding the ploidy level of the *Rosa* genome when BLAST-searched on already assembled *Rosa* datasets. The final set contains 1,856 tags that could be used for phylogenomic analyses (Additional file 1). These tags will now be referred to as Single-Copy Orthologous Tags ($\mathrm{SCO_{Tag}}$s) in the text, to denote that they are short, PCR-amplifiable sequence tags, derived from primers in conserved sequences that flank variable sequence regions in single-copy orthologous genes identified across a set of closely-related species. Of these 1856 $\mathrm{SCO_{Tag}}$s, 1223 (66%) cover both coding and non-coding regions, while 550 (30%) cover pure coding regions and 83 (4%) cover pure non-coding regions (Figure 26C).

We also searched outgroup species genomes for the presence of the respective 1856 $\mathrm{SCO_{Tag}}$s, leading to 1534 $\mathrm{SCO_{Tag}}$s that contain at least one of the seven outgroup species (*Fragaria iinumae*: 1029; *F. nipponica*: 875; *F. nubicola*: 858; *F. vesca*: 1142; *Rubus occidentalis*: 985; *Geum urbanum*: 697; *Potentilla micrantha*: 1092). Apart from *Rosa* 'Old Blush' and *R. persica*, which are present for all of the 1856 $\mathrm{SCO_{Tag}}$s, the taxon occupancy of $\mathrm{SCO_{Tag}}$s for the *Rosa* ingroup varies from 23% for *R. palustris* to 97% for *R. wichurana* (Appendix A, Supplementary figure A.3). Half of the 1856 $\mathrm{SCO_{Tag}}$s have been found in at least 14 out of the 17 *Rosa* species analyzed. Species sequences from each of the 1856 $\mathrm{SCO_{Tag}}$s are available in Additional file 2. Species sequences from each $\mathrm{SCO_{Tag}}$ were aligned using mafft and cleaned with Gblocks, leading to a supermatrix of 669,354 bp for the ingroup species with 28% of missing data, after the removal of 4843 (0.7%) poorly-aligned sites. For the dataset with ingroup plus outgroup species, the supermatrix contained 676,389 bp with 34% of missing data after the removal of 16,978 (2.4%) poorly-aligned sites.

2.4.3 Efficiency of plastid loci and nuclear $\mathrm{SCO_{Tag}}$s for *Rosa* phylogeny

We analyzed the sequence variation contained in each of the 1856 $\mathrm{SCO_{Tag}}$ alignments, focusing only on the *Rosa* ingroup. The mean number of taxa per $\mathrm{SCO_{Tag}}$ alignment was 9, 11 and 15 for $\mathrm{SCO_{Tag}}$s covering non-coding, mixed and coding regions, respectively. As expected, on average, the non-coding regions contain more parsimony-informative sites (PIS) than mixed sequences, which in turn contain more PIS than pure coding regions (Figure 26D). Plastid sequences trnL and trnG have medium PIS content (2-3%), whereas the psbA-trnH region is highly variable (>8% of PIS) and reaches the upper bound of PIS content distributions of both mixed and non-coding sequences (Figure 26D).

In the nuclear $\mathrm{SCO_{Tag}}$ species-chronogram, almost all branches show bootstrap supports (BS) of 100%, in clear contrast with the species-tree obtained based on the conventional plastid sequences (Figure 27). Both datasets support a distinct *Chinenses*-*Gallicanae*-*Synstylae* clade but have slightly different tree structures for the remaining species. While only the nuclear $\mathrm{SCO_{Tag}}$s support monophyly for the *Chinenses* and three of the four *Synstylae*, both datasets exhibit strong support (>99% BS) for the position of *Rosa moschata* and *R. minutifolia* near the *Rosa* clade. In addition, the nuclear $\mathrm{SCO_{Tag}}$s dates the *R. laevigata* speciation event as being more ancient (26 MYa) than the plastid dataset suggests (16 MYa) and supports the monophyly of the two bright yellow-flowered species, *R. persica* and *R. xanthina*.

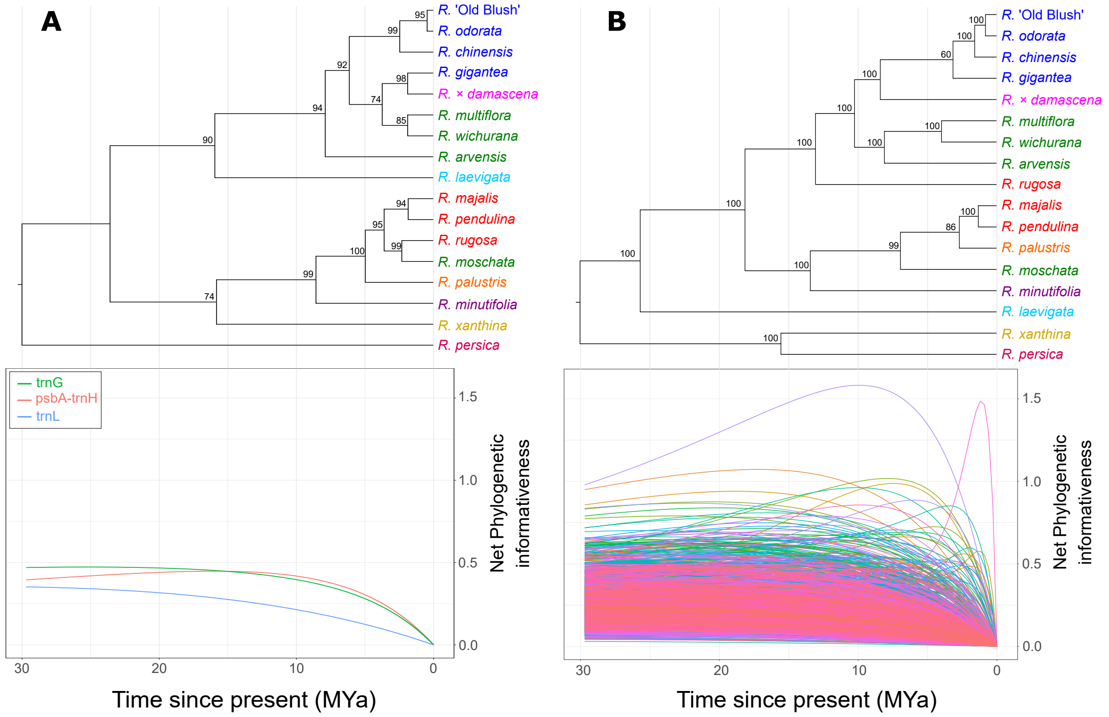


Figure 27: Net phylogenetic informativeness (PI) profiles compared to species chronograms. A) Plastid loci; B) 1856 nuclear $\mathrm{SCO_{Tag}}$s. Taxa are colored as follows: dark blue for taxa from *Rosa* sect. *Chinenses*, pink for *R.* sect. *Gallicanae*, green for *R.* sect. *Synstylae*, light blue for *R.* sect. *Laevigatae*, red for *R.* sect. *Rosa* (ex. *R.* sect. *Cinnamomeae*), orange for *R.* sect. *Carolinae*, purple for *R.* subg. *Hesperhodos*, yellow for *R.* sect. *Pimpinellifoliae* and fuchsia for *R.* subg. *Hulthemia*.

2.4.4 Phylogenetic Informativeness

The Phylogenetic Informativeness (PI) profiles of the plastid sequences are smooth, with a slow decrease through geological time, and they never reach values above net PI of 0.5 (Figure 27A). During the last 8 M years, psbA-trnH and trnG display a similar profile but trnG reaches higher PI values for more ancient periods. The trnL locus shows lower PI values than the two other plastid loci at all times. The PI profiles of the nuclear $\mathrm{SCO_{Tag}}$s have different shapes and heights (Figure 27B). While most of the $\mathrm{SCO_{Tag}}$s do not exceed a net PI of 0.5 during the past 30 M years of divergence, some reach PI values higher than 1.0. A total of 131 $\mathrm{SCO_{Tag}}$s reach their maximum value at the 0-15 MYa time interval, which represents the most recent half of the total divergence period and includes 75% of the species-tree nodes. For older nodes, informative $\mathrm{SCO_{Tag}}$s can be identified with PI values peaking around 20 MYa with net PI between 0.75 and 1. Additionally, we observed that the area under the PI profiles for the time interval 0-30 MYa tends to decrease while more taxa are added to $\mathrm{SCO_{Tag}}$ alignments ($y = 11.8 - 0.45x, R^2 = 0.18$). By increasing the number of taxa per alignment from 6 to 17, the average area under the PI profile decreases by a factor of 2 (Appendix A, Supplementary figure A.4A). Albeit less clear, the fraction of variable sites in $\mathrm{SCO_{Tag}}$ alignments also tends to be negatively correlated with the number of taxa included per $\mathrm{SCO_{Tag}}$ alignment, especially for $\mathrm{SCO_{Tag}}$ with high taxon occupancy ($y = 22.1 - 0.90x, R^2 = 0.11$) (Appendix A, Supplementary figure A.4B).

2.4.5 Analysis of topological conflict

Higher PI profiles of nuclear $\mathrm{SCO_{Tag}}$s at a time interval do not necessary correspond to better support values in the corresponding species-chronogram. This is because PI does not directly account for phylogenetic noise (Townsend, 2007), so that genes with fast-evolving sites may display high PI profiles, whereas they can increase the number of homoplastic sites and obscure the number of synapomorphic sites which therefore scrambles the phylogenetic signal and provides poor support for bipartitions (Klopfstein et al., 2010). Therefore, we also tested our $\mathrm{SCO_{Tag}}$s based on topological criteria to ensure that highly informative $\mathrm{SCO_{Tag}}$s are concordant with the species-tree and do not result from regions with too many fast evolving sites. We first constructed a network to summarize conflicts between all $\mathrm{SCO_{Tag}}$s trees (Appendix A, Supplementary figure A.5). Species groups identified in the network are mostly consistent with the clades found in the concatenated analysis (Figure 27B). The reticulation pattern show conflict between $\mathrm{SCO_{Tag}}$ trees for both recent and ancient speciations. For recent speciations, links between species are short and packed while they are long and slack for more ancient speciations. Then, we detailed these conflicts for each node of the species-tree using PhyParts. Of the 1534 $\mathrm{SCO_{Tag}}$s with at least one outgroup sequence, 8 did not resolve the monophyly of outgroup species and were therefore discarded since rooted $\mathrm{SCO_{Tag}}$ trees are required to detail the underlying conflict at each node of the species-tree. The Maximum Likelihood (ML) species-tree obtained after concatenation of the 1526 resulting $\mathrm{SCO_{Tag}}$s is presented in Figure 28. The topology is the same as for (1) the coalescent species-tree obtained after the reconciliation of the 1,526 $\mathrm{SCO_{Tag}}$ trees and (2) the chronogram presented in Figure 27B, but with slight modification of BS for node 7 (increase from 60% to 75%), node 13 (decrease from 99% to 91%), node 14 (decrease from 86% to 65%), node 16 (decrease from 100% to 79%). In addition to BS, we computed two other support values: (1) Local Posterior Probabilities (LPP) that derive from frequencies of quadripartitions observed in the set of $\mathrm{SCO_{Tag}}$ trees and (2) Internode Certainty All (ICA) scores that provide information on the amount of conflict at each node. Although not directly related, these three support values each explain in their own way the phylogenetic signal present in the dataset. We observe that low LPP generally correspond to less supported branches (BS $<$ 100%), except for node 16. However, we often observe that high LPP and BS value do not always correspond to high ICA scores (Node 6, 8, 9 and10). The normalized quartet score for the coalescent tree is 0.73, meaning that 73% of all the quadripartitions found in $\mathrm{SCO_{Tag}}$ trees satisfy the coalescent species-tree. We then deconstructed each $\mathrm{SCO_{Tag}}$ tree topology and focused only on bipartitions showing >70% BS that we compared to the bipartitions found in the ML species tree. $\mathrm{SCO_{Tag}}$s resolve more bipartitions with a BS >70% at ancient nodes than at recent nodes. This observation holds as well for the ICA score where the most ancient nodes have higher ICA values than the most recent nodes (Figure 28). For very recent nodes, few $\mathrm{SCO_{Tag}}$s can individually make the distinction between closely related taxa at this BS threshold.

Regarding patterns of concordance and conflict, we first observe that no $\mathrm{SCO_{Tag}}$s are concordant with more than six of the 16 nodes present in the species-tree (Figure 29A), whereas some $\mathrm{SCO_{Tag}}$s are in conflict with up to 12 nodes (Figure 29B). The highly conflicting $\mathrm{SCO_{Tag}}$s (conflicting in more than seven nodes) represent a minority (4%) of the entire dataset. Actually, 625 $\mathrm{SCO_{Tag}}$s bear 0 conflicting nodes and 1184 $\mathrm{SCO_{Tag}}$s agree with one to three nodes. Then, we analyzed the pattern of conflict node by node. We observed that more than two-thirds of the $\mathrm{SCO_{Tag}}$s agree in dividing the genus at node 1 with the two yellow-flowered species *Rosa persica* and *R. xanthina* separate from the rest of the *Rosa* species. For more recent nodes, higher number of individual alternative bipartitions can be observed (Figure 28). Nodes 7, 9, 10, 13, 14 and 16 show a significant proportion of $\mathrm{SCO_{Tag}}$s agreeing with the main alternative bipartition, meaning that the proportion of $\mathrm{SCO_{Tag}}$s supporting the main alternative bipartition is greater than 50% of the proportion of $\mathrm{SCO_{Tag}}$s agreeing with the species-tree bipartition (Appendix A, Supplementary figure A.6). These conflicting nodes do not always correspond to the lowest BS, ICA or LPP support values.

2.4.6 Correlation between phylogenetic informativeness and topological conflict

We then correlated the area under the PI profile for the 0-30 MYa time interval with the number of nodes in $\mathrm{SCO_{Tag}}$ tree that are concordant or in conflict with the species-tree, using a BS cutoff of 70% (Figure 29). We observed that PI tends to increase while more concordant nodes are present in $\mathrm{SCO_{Tag}}$ trees ($y = 4.95 + 0.65x, R^2 = 0.04$). A similar observation can be made for the number of conflicting nodes ($y = 5.03 + 0.56x, R^2 = 0.10$). Interestingly, we observed that the

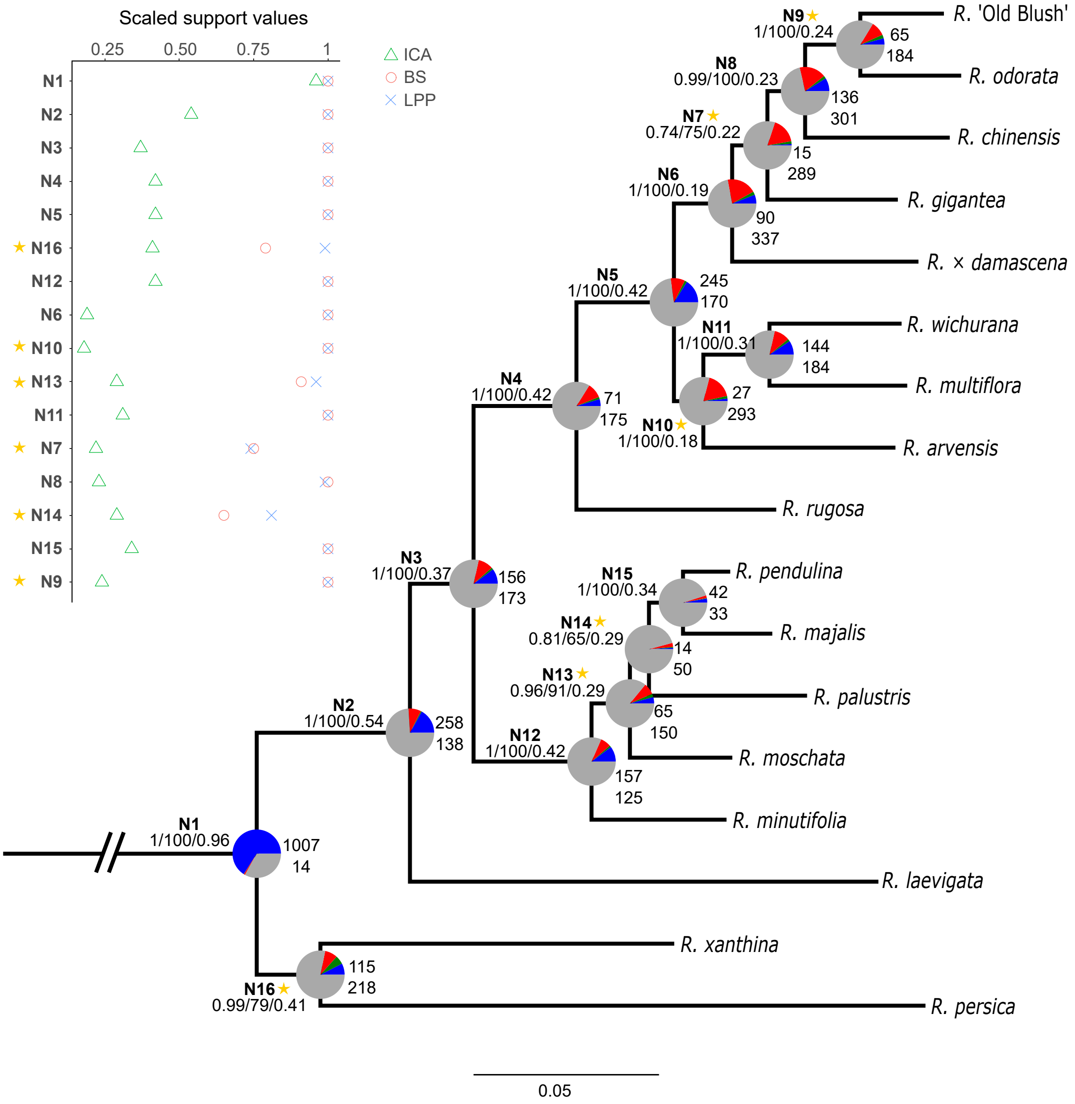


Figure 28: Combined ML species tree with summary of conflicting and concordant SCO$_{\text{Tag}}$s. The ML species-tree was constructed from 1526 concatenated rooted SCO$_{\text{Tag}}$s. Outgroups are not shown. Node names are in bold. For each branch, the three values separated by a slash are the local posterior probability (LPP), the bootstrap support (BS) and the Internode Certainty All (ICA), respectively. The pie charts at each node present the fraction of SCO$_{\text{Tag}}$s that supports that bipartition (blue), the fraction that supports the main alternative bipartition (green), the fraction that supports other alternative bipartitions (red) and the fraction with either less than 70% BS at this bipartition or that do not have this partition due to missing data (gray). On the right side of the pie charts, the top and bottom values indicate the numbers of SCO$_{\text{Tag}}$s concordant, respectively in conflict, with the corresponding bipartition in the species-tree. Scatter plot on the left side compares values of BS, LPP and ICA at each node. Nodes are ranked from the most ancient (N1) to the most recent (N9) according to Figure 28B. Stars indicate conflicting nodes with great fractions of alternative bipartitions.

top most informative SCO_{Tag} identified in Figure 27B is in fact conflicting in 10 nodes and agrees with 0 node (Figure 29). In addition, the 330 SCO_{Tag}s that have not been analyzed for topological concordance due to lack of outgroups tend to show a similar PI distribution to SCO_{Tag}s that were analyzed for topological conflict (Figure 29C).

Metrics regarding variability content, phylogenetic informativeness and topological conflict for the 1856 SCO_{Tag}s are available in Additional file 3.

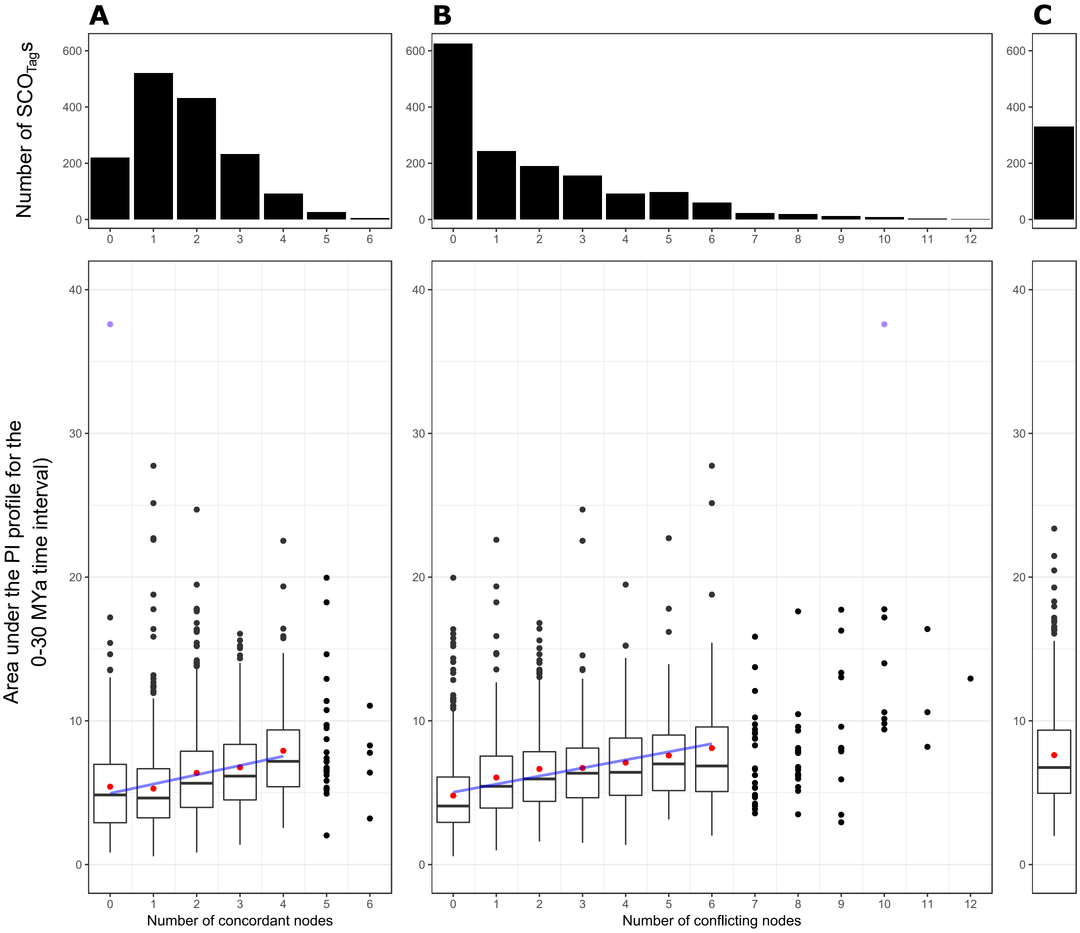


Figure 29: Correlation between phylogenetic informativeness (PI) and the number of **A**) concordant nodes and **B**) conflicting nodes in SCO_{Tag} topologies. **C**) corresponds to the PI distribution for unrootable SCO_{Tag} that were not analyzed using PhyParts. Situations with less than 30 points were ploted but not used in the calculation of correlations. Red dots correspond to mean values. Blue lines correspond to regression lines: $y = 4.95 + 0.65x, R^2 = 0.04$ in panel A and $y = 5.03 + 0.56x, R^2 = 0.10$ in panel B. The top most purple dot corresponds to the highest PI profile in Figure 27.

2.5 Discussion

2.5.1 Finding nuclear $\mathrm{SCO_{Tag}}$s at the genus level

Several sets of SCOs have recently been released but few studies have focused on developing SCOs dedicated to species-level phylogeny (Granados Mendoza et al., 2015; Kates et al., 2017). For genera such as *Rosa*, which shows rapid radiations (Joly et al., 2006b; Herklotz and Ritz, 2017), it is likely that DNA sequences (either nuclear or plastid) are very closely-related, and SCOs designed for reconstructing the broad angiosperm phylogeny may not be suited to resolving species relationships. In this study, we took the woodland strawberry (*Fragaria vesca*) as an outgroup to identify SCOs shared with the genus *Rosa*. The two taxa share similar genome characteristics such as diploidy and a base chromosome number of seven. Macro-synteny analysis also revealed only one major translocation event between two chromosomes (Hibrand Saint-Oyant et al., 2018). In addition, *Fragaria vesca* and *Rosa* species belong to sister tribes within the subfamily Rosoideae (Xiang et al., 2016). The number of single-copy genes that we identified in each of the two species was consistent with previous observations across angiosperms (Han et al., 2014). *Rosa* 'Old Blush' is currently the only *Rosa* taxon with a high-quality annotated genome sequence and we chose it as the reference for the whole *Rosa* genus (Hibrand Saint-Oyant et al., 2018). We identified 1784 conserved genes in the subfamily Rosoideae by searching for shared SCOs between *Fragaria vesca* and *Rosa* 'Old Blush'. We observed a relative shared synteny in the localization of the 1784 SCOs between *F. vesca* and *R.* 'Old Blush' which emphasizes on the fact that we selected conserved genes. We also found that 73% of the 1784 SCOs are not present in other published ortholog sets (Appendix A, Supplementary figure A.2), suggesting that it is worth developing specific phylogenomic markers that are dedicated to each particular taxonomic group. Then, we considered that the 1784 SCOs identified in *R.* 'Old Blush' are also orthologous in other *Rosa* species. We therefore assumed that no more gene duplication or gene loss occurred in the SCO set after the divergence of the *Potentillae* and the *Roseae* tribes around 60 MYa (Xiang et al., 2016). No recent large genome duplication was detected in *Rosa* 'Old Blush' (Hibrand Saint-Oyant et al., 2018). Since *Rosa* 'Old Blush' is considered to be an interspecific hybrid between *R. odorata* and *R. chinensis* (Meng et al., 2011), two species sharing their last common ancestor some 8-9 MYa (Fougère-Danezan et al., 2015), this suggests that gene gain by large duplication is not common in closely-related roses. However, our assumption may not hold if fine-scale genome rearrangements occurred in other *Rosa* species that were not analyzed here. This means that paralogous genes might be targeted using our 1784 SCOs on a broader set of *Rosa* species. For this reason, we carried out additional filtering on the tags obtained after the target assembly of the 1784 SCOs. This filtering procedure aimed to eliminate putative paralogous sequences by discarding (1) tags with unspecific primer pairs (Step 5a, Figure 25), (2) tags that do not have a strict 1-to-1 orthologous relationship with the reference genome of *R.* 'Old Blush' (Step 5b, Figure 25) and (3) tags with an inconsistent number of alleles in either aTRAM contigs (Step 5c, Figure 25) or already assembled *Rosa* genomes (Step 5d, Figure 25). Our final set of 1856 SCO_{Tag}s derived from 1784 SCOs should therefore essentially contain orthologous sequences suited to phylogenomics analyses.

Using shotgun sequencing libraries and Illumina short-read sequencing at low depth (10-30x) in 12 *Rosa* species, we applied the aTRAM pipeline to assemble specific loci (Allen et al., 2015), and we retrieved most of the 1784 SCOs (Figure 26B). While this method does not take individual heterozygosity at each SCO_{Tag} into account, it provides a fast and easy way to extract genome sequences of specific loci, while circumventing whole genome assemblies, which may be particularly difficult for highly heterozygous taxa such as *Rosa* species. Our procedure only retains one representative SCO_{Tag} sequence per species, which may be sufficient for genus section comparisons. However, phylogenomic analyses below the section level may require to reconstruct multiple sequence variants per species to reveal hybrid specimens. For this reason, we developed conserved SCO_{Tag} primer pairs that can be used to target SCO_{Tag} alleles using basic PCR amplifications in future analyses (Additional file 1).

2.5.2 Efficiency of nuclear SCO_{Tag}s for phylogenomics in the genus *Rosa*

We have built a ML phylogenomic tree of some representative species of the genus *Rosa* using 1856 SCO_{Tag}s within 1784 SCOs (Figure 27B), leading to a highly supported species-tree structure. Both plastid loci and nuclear SCO_{Tag}s revealed a *Chinenses*-*Gallicanae*-*Synstylae* clade, but only the nuclear SCO_{Tag}s supports the monophyly of these three groups within the clade. On the contrary, the plastid loci better resolve the monophyly of the *Rosa* sect. *Rosa* clade, while *R. rugosa* is separated from the rest of *Rosa* sect. *Rosa* species, and found near the *Chinenses*-*Gallicanae*-*Synstylae* clade in the nuclear SCO_{Tag} topology. Compared to previous studies (Bruneau et al., 2007; Fougère-Danezan et al., 2015), both plastid and nuclear sequences expressed unexpected positions of *R. moschata*, which was expected to group together with the other *Synstylae*, and *R. minutifolia* that was expected to branch off earlier in the phylogenetic tree. These discrepancies may arise from the taxon sampling itself. *R. moschata* and *R. rugosa* have been extensively used in breeding (Wylie, 1954) and Hibrand Saint-Oyant et al. (2018) may have sampled one of many varieties that were derived from hybridization. The wild origin of *Rosa moschata* is uncertain (Masure, 2013) since several moschata-type roses share a similar geographical distribution from Southeast Europe to the Himalayas, such as *R. beggeriana* Schrenk ex Fisch. & C. A. Meyer, *R. fedtschenkoana* Regel and *R. brunonii* Lindl. (Schramm, 2016). The latter is often cultivated as *R. moschata* in rose gardens (Rehder, 1940). We suggest that the *R. moschata* that we used could be a hybrid between several wild species sharing a common distribution, with at least one species (*R. beggeriana*) belonging to *R.* sect. *Rosa*, the same section as *R. rugosa*. This could explain that *R. moschata* is closely related to the *R.* section *Rosa* in our analysis (Figure 27). Based on a comparison between plastid loci and nuclear SCO_{Tag}s phylogenies, our data may suggest that the maternal origin of our *R. rugosa* is from *R.* sect. *Rosa*, whereas its nuclear genome shows proximity with species of *R.* sect. *Synstylae*, also native to Northeast Asia. This demonstrates the utility of combining plastid and nuclear sequences for phylogenomic analyses to reveal putative hybridization events. The *R. minutifolia* we analyzed here is a white variety known as *R. minutifolia* 'Alba', and the accession used shows unexpected morphological characteristics (leaflet size >3 cm, long pinnate leaves and multi-flowered inflorescences), suggesting an earlier cross with a species from *Rosa* subg. *Rosa*. The ease for *Rosa* species to hybridize poses a major challenge for correct taxonomic identification. This highlights the importance for future studies to preferentially sample several specimens per species, including wild accessions and garden-grown accessions derived from cuttings with a known wild origin.

We further evaluated which of the 1856 $\mathrm{SCO_{Tag}}$s performed best for a future phylogenomics study on a broader set of wild species in the *Rosa* genus. PI analyses showed that a large fraction of nuclear $\mathrm{SCO_{Tag}}$s have little information content to reconstruct speciation events in the genus *Rosa* with profiles lower than 0.5 of net PI and a slow decrease over time (Figure 27A). However, a few hundred $\mathrm{SCO_{Tag}}$s exhibit high PI profiles that peaked at different ages of the chronogram. Such a diversity of PI profiles is interesting since different sets of $\mathrm{SCO_{Tag}}$s could resolve specific levels of the species-tree. Many of the ancient nodes are not well supported in recently published plastid phylogenies of the genus *Rosa* (Bruneau et al., 2007; Fougère-Danezan et al., 2015) and it would be interesting to target $\mathrm{SCO_{Tag}}$s with high PI during ancient evolutionary time intervals. PI profiles of conventional plastid sequences show their limitations to resolve nodes in *Rosa* phylogeny, even for psbA-trnH (Figure 27A) that has a relatively high PIS content, in line with previous works that compared phylogenetic informativeness of nuclear vs. plastid sequences in other groups (Granados Mendoza et al., 2015).

We then focused on topological conflict between each $\mathrm{SCO_{Tag}}$ tree toward the species-tree (Figure 28). We mainly show that most $\mathrm{SCO_{Tag}}$s cannot individually resolve shallow to intermediate nodes with a BS threshold of 70%. One of the main reasons may be the alignment length of each $\mathrm{SCO_{Tag}}$ which is very short and barely exceeds 500 bp. It may therefore be difficult to have enough variable sites for a good confidence in bipartitions within only one $\mathrm{SCO_{Tag}}$, especially for recent times where DNA sequences among closely-related taxa are expected to be very similar. $\mathrm{SCO_{Tag}}$s that display bipartitions with a BS >70% often support alternative bipartitions that do not reflect the species-tree. These discrepancies between gene-trees and the species-tree were already observed in other studies (Maddison and Wiens, 1997; Nichols, 2001). Global patterns of conflict were first summarized on a network (Appendix A, Supplementary figure A.5) and further detailed node by node. We observed that the species network highlighted many conflicts between $\mathrm{SCO_{Tag}}$ trees although the species groups identified were consistent with the ML species tree. Recent divergences were more prone to conflict as observed with the tight links between close-related species on the network and further confirmed by the decrease of ICA scores for recent nodes (Figure 28). In details, several nodes showed a high proportion of the main alternative bipartition (Appendix A, Supplementary figure A.6). Most of them concern rearrangements between species inside a section clade or between neighboring species that belong to sister clades in Figure 27B. Conflicts observed at node 7 and node 9 relate to switches between species that belong to the *Chinenses-Gallicanae* clade. For instance, the main alternative bipartition found for node 9 involves the switches between

Rosa odorata, *R. gigantea* and *R. chinensis* as the species that are the most closely related to *Rosa* 'Old Blush'. Those tree structures can be explained since *R. chinensis* and *R. odorata* var. *gigantea* are probably the parents of *Rosa* 'Old Blush' (Meng et al., 2011). The reference genome sequence of *Rosa* 'Old Blush' was obtained from a haploid cell line derived from pollen cells (Hibrand Saint-Oyant et al., 2018). The resulting chromosome set may have contained unequal contributions from the ancestral *R. odorata* and *R. chinensis* genomes after the random meiotic division. Conflicts at node 10 comes from the switch between *Synstylae* species and *Chinenses-Gallicanae* species, showing the close relationships between those sections. The dubious positioning of *R. minutifolia* brings conflicts at node 13 and 14 since *R. minutifolia* is found sometimes closer to *R. pendulina* (*R.* sect. *Rosa*), sometimes closer to *R. moschata* (*R.* sect. *Synstylae*), highlighting again the issue of correct taxonomic identification of this accession. Finally, the most ancient node with a significant main alternative bipartition is node 16 and relates to the split of the clade *R. xanthina*, *R. persica* into two separate lineages. Despite their bright yellow petals, *R. persica* and *R. xanthina* are very different wild rose species in terms of shapes, habitats and morphological traits (Rehder, 1940; Masure, 2013). Sampling additional rose species in *Rosa* sect. *Pimpinellifoliae* will be useful in future studies to resolve how these species are related.

2.5.3 Impact of missing data and topological conflict in $\mathrm{SCO_{Tag}}$s selection

In this study, we had to deal with missing or partial data for almost all of the 1784 SCOs (Figure 26B) and therefore for almost all of the 1856 resulting $\mathrm{SCO_{Tag}}$s (Appendix A, Supplementary Figure A.3). Since the approach to $\mathrm{SCO_{Tag}}$ identification involves primer design in strictly conserved sequences flanking variable regions, we only kept SCO alignments covered by at least four taxa, including the reference genome sequence of *Rosa* 'Old Blush' and the highly divergent species *Rosa persica*. The variation in the number of species included in the 1784 respective SCO alignments has several underlying reasons and has associated consequences for downstream analysis. The underlying reasons for missing species from SCO alignments may reflect: (1) the actual gene duplication or gene loss in the genome of a given species; (2) insufficient read depth or inability to reconstruct the locus from the whole genome shotgun sequencing data; (3) strong sequence divergence that hampers the recognition of high confidence BLAST identification of orthologous genes from a given species. Furthermore, selecting informative $\mathrm{SCO_{Tag}}$s depends on the complex relationship between the number of taxa compared, their sequence divergence (which, in turn, depends on coding/non-coding capacity) and parsimony-informative site (PIS) content. For instance, the more taxa that are compared and the more divergent the species that are included in the alignment are, the more likely it is that variable sites will become parsimony informative, but the less likely it is to identify flanking, strictly conserved regions for primer design. Indeed, classification of the coding potential of $\mathrm{SCO_{Tag}}$s based on positional overlap with structural gene model annotation revealed, as expected, that non-coding $\mathrm{SCO_{Tag}}$ alignments comprise two-fold less species than pure coding $\mathrm{SCO_{Tag}}$ alignments, in line with elevated sequence divergence in non-coding regions compared to protein coding sequences. As a consequence, SCO_{Tag}s that contain strictly non-coding regions comprise only 4% of the entire SCO_{Tag} set, and while they contain lower numbers of taxa per alignment, they still exhibit the highest relative PIS content (Figure 26D). A substantial fraction of our SCO_{Tag}s contains both coding and non-coding regions, and selecting this type of SCO_{Tag} may be a good strategy to target conserved regions surrounding variable sequences. By increasing the relative fraction of non-coding SCO_{Tag}s, the procedure proposed here may be more informative than exon capture or phylotranscriptomics to decipher phylogenetic relationships for closely-related species or those with complex evolutionary relationships.

Furthermore, we analyzed our set of SCO_{Tag}s for phylogenomic informativeness as a function of divergence time as well as for topological conflict. We observed lower PI values for SCO_{Tag}s containing the most taxa (Appendix A, Supplementary Figure A.4A), suggesting that well-covered SCO_{Tag}s would not be preferentially sampled based on the PI profile criteria. Klopfstein et al. (2010) claim that adding more taxa to the alignment reduces the probability to observe a never-reversed synapomorphy since each new taxon may reverse the synapomorphy and thus lower the optimum evolutionary rate. In contrast, Townsend and Leuenberger (2011) argued that increasing taxon sampling does not decrease that optimal rate of character change. Here, all SCO_{Tag} alignments contain sequences of the most divergent wild rose species and at least two other intermediate species. It is therefore unlikely that some loci disproportionally represent ancient vs. recent divergences. We also observed that SCO_{Tag}s with few taxa tend to have greater relative numbers of variable sites (Appendix A, Supplementary figure A.4B), which may be due to the fact that SCO_{Tag}s with less taxon occupancy are less conserved and therefore more variable.

Townsend's PI does not directly account for noise that may be caused by fast-evolving sites. However, a thorough analysis of PI curves can provide insight into how much noise is present in each SCO_{Tag}. Sharp recent peaks with a steep post-slope may introduce noise for older nodes. Consequently, for a given value of PImax, it is better to select SCO_{Tag}s that express a steady decline after they peak (Townsend and Leuenberger, 2011; Hilu et al., 2014). Despite we did not observed a general strong correlation between PI and topological conflict, we noticed that the top most informative SCO_{Tag} for the 0-30 MYa time interval (Figure 27B) is also a highly conflicting SCO_{Tag} (Figure 29B). This demonstrates the importance to combine different approaches to evaluate the set of sequences prior to phylogenomics inferences. This assessment enables to identify the most phylogenetic informative sequences and to reveal patterns of conflicts while a basic supermatrix approach simply conceals conflicts and can even produces a well-supported but incorrect species tree (Kubatko et al., 2009; Salichos and Rokas, 2013). Atypical SCO_{Tag} should not necessary be removed for downstream phylogenomic analyses since they hold different evolutionary histories that may be interesting to study. Regarding phylogenomics in the genus *Rosa*, the many patterns of conflict, that we especially observed in close-related species, highlight the difficulty to clearly identify one overall evolutionary history in this genus. Patterns of conflicts will have to be taken into account in future studies to accurately unravel the complex mechanisms that shaped this genus. It is also worth mentioning that our sampling covers only one-tenth of the existing wild rose species, and some recent rapidly evolving sections such as *Rosa* sect. *Caninae* are not represented. Thus, we recommend selecting well-covered $\mathrm{SCO_{Tag}}$s, that peak at various times during the 30 M years of divergence for future studies on *Rosa* relationships. Using sets of $\mathrm{SCO_{Tag}}$s with similar PI values, $\mathrm{SCO_{Tag}}$s with maximal numbers of species should be prioritized to increase the chance of successful target PCR amplification.

2.6 Conclusions

The method implemented here to mine genome-scale sequencing data successfully recovered hundreds of nuclear single-copy orthologous sequence tags suitable for species-level phylogenomics in the highly complex genus *Rosa*. We emphasize that a thorough analysis must be performed on phylogenomic datasets in order to choose the most informative markers. While the sequence content of variable sites is obviously important, it does not predict better topology resolution. Computing phylogenetic informativeness and topological conflict of $\mathrm{SCO_{Tag}}$s ensures the selection of a comprehensive set of $\mathrm{SCO_{Tag}}$s containing appropriate sequence variations to cover the entire period of species divergence and simultaneously reveals potential sources of topological conflict that may have biological meanings, such as hybridization events or unwanted selection of paralogous copies. Despite the fact that plastid sequences are less variable, their one-sided inheritance still gives valuable perspectives for comparison with nuclear data in view of a better understanding of how evolutionary processes, such as hybridization, shape complex genera such as *Rosa*. The mining strategy presented here enables the development of $\mathrm{SCO_{Tag}}$ nuclear markers to target yet unresolved parts of the green plants' Tree of Life, from the deepest branches to the shallowest relationships between individuals.

2.7 Additional file and availability of data and material

Additional File 1. List of the 1856 $\mathrm{SCO_{Tag}}$s primer pairs for *Rosa* phylogenomics (Excel file). This file list the primer sequences associated with the 1856 $\mathrm{SCO_{Tag}}$s, as well as information about melting temperature and corresponding $\mathrm{SCO_{Tag}}$ amplicons in the haploid reference genome sequence of *Rosa* 'Old Blush' (Hibrand Saint-Oyant et al., 2018) (genome coordinates and fragment length).

Additional File 2. Sequences of the 1856 across seven Rosaceae outgroups and 12 *Rosa* species (Fasta file). Raw fasta sequences associated with the 1856 $\mathrm{SCO_{Tag}}$s per species. These sequences correspond to either target-assembled $\mathrm{SCO_{Tag}}$s from whole genome shotgun Illumina paired-end reads or $\mathrm{SCO_{Tag}}$s that were found in already assembled datasets.

Additional file 3. Metrics associated with the 1856 $\mathrm{SCO_{Tag}}$s (Excel file). This file contains all metrics that served to assess the phylogenetic utility of the 1856 $\mathrm{SCO_{Tag}}$s. Metrics such as sequence variability, phylogenetic informativeness, node-by-node topological conflict and structural annotation are detailed for each of the 1856 $\mathrm{SCO_{Tag}}$s.

All the above-mentioned additional file are available as part of the supplementary data associated with Debray et al. (2019) which article can be found at:

https://bmcevolbiol.biomedcentral.com/articles/10.1186/s12862-019-1479-z

2.8 Author's contributions

KD, FF, VM conceived and designed the study. KD performed the mining analysis and drafted the manuscript. FF, VM, JC and TR proof read the manuscript. JMM helped with bioinformatics and retrieving plastid sequences. TR provided early access to the *Rosa wichurana* genome assembly. JC provided early access to the sequencing data published in Hibrand Saint-Oyant et al. (2018). All authors have read and approved the manuscript.

2.9 Acknowledgments

The authors are grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrenées (Bioinfo GenoToul) for providing computing and storage resources and bioinformatics support. The authors also thank Gail Wagman for reviewing the English and two anonymous reviewers for valuable remarks on the manuscript.

2.10 Supplementary results: On the origin of *Rosa chinensis* 'Old Blush'

2.10.1 Introduction

The previous chapter developed a general method to identify informative nuclear single-copy tags ($\mathrm{SCO_{Tag}}$s) at a specific taxonomic level. The genus *Rosa* was taken as an example and the pipeline yielded a total of 1856 $\mathrm{SCO_{Tag}}$s, in sharp contrast with the few nuclear sequences or so that were commonly used in *Rosa* studies until present. With this set of $\mathrm{SCO_{Tag}}$s, the possibilities to study the nuclear genomes of *Rosa* species are therefore broadened. $\mathrm{SCO_{Tag}}$s have indeed many advantages for comparing the genome evolution in a set of species. First, $\mathrm{SCO_{Tag}}$s were designed with specific conserved primer pairs at 5' and 3' ends, making them easily amplifiable using the cheap and common PCR technique, with the possibility to multiplex. Then, conserved primer pairs flank more variable regions which phylogenetic informativeness was appraised at different epochs of the 30 MY of divergence of the genus. In addition, the orthology of $\mathrm{SCO_{Tag}}$s has been assessed by selecting single-copy genes and performing reciprocal best BLAST searches. $\mathrm{SCO_{Tag}}$s therefore meet all the criteria to be good phylogenetic markers. Finally, $\mathrm{SCO_{Tag}}$s are well-distributed along the seven pseudomolecules of the reference genome and can therefore serve to compare the genomic organization between *Rosa* while avoiding the expensive whole genome sequencing steps and the burden associated with genome assemblies. SCO$_{\text{Tag}}$s can also be used to study synteny with related taxa, or to anchor scaffolds from fragmented *Rosa* assemblies.

Rosa 'Old Blush' is an ancient diploid rose also known as 'Parson's Pink China' or 'Old Blush China'. It originates from China and was introduced in Europe during the 18$^{\text{th}}$ century, along with three other double-petaled cultivars: 'Hume's Blush Tea-scented China', 'Parks' Yellow Tea-scented China' and 'Slater's Crimson China' (Joyaux, 2015). *R.* 'Old Blush' was appreciated for its scent and recurrent blooming that were not found in European-Mediterranean rose cultivars at that time. The origin of *R.* 'Old Blush' is much debated and remain only putative mainly because it has been in cultivation for centuries and that there exist a large number of synonyms for this cultivar. Considering the putative wild progenitors of *R.* 'Old Blush', several authors reported that it may originate from a cross between *R. chinensis* Jacq. and *R. gigantea* Collet (*R.* × *odorata* var. *gigantea* (Collet ex Crép.) Rehd. & Wils. 1915) (Hurst, 1941; Wylie, 1954; Joyaux, 2015). *R.* 'Old Blush' is often considered as a variety of *Rosa chinensis*, although Meng et al. (2011) categorized *R.* 'Old Blush' (syn. 'Parson's Pink China') as *R. odorata* var. *erubescens* based on a close reading of Hurst's (1941) descriptions. *R.* 'Old Blush' can also be found under the name *R.* × *odorata* 'Pallida', although the phylogeny presented in Qiu et al. (2013) clearly separate *R. odorata* var. *erubescens* and 'Pallida' in distinct clades. Using molecular data, Meng et al. (2011) further confirm that the three cultivars of *R. odorata* ('Hume's Blush Tea-scented China', 'Parson's Pink China' [i.e. = 'Old Blush'] and 'Park's Yellow Tea-scented China') introduced in Europe at the turn of the 18th century are a mix between *R. odorata* var. *gigantea* and different local cultivars of *R. chinensis*. Based on the analysis of plastid sequences, possibly maternally inherited in *Rosa* (Corriveau and Coleman, 1988), Meng et al. (2011) suggested that *R. odorata* var. *gigantea* could be the maternal parent of the three double-petaled cultivars, in contrast with observations made in this chapter and in other studies (Tan et al., 2017; Raymond et al., 2018). Here, we used SCO$_{\text{Tag}}$s for painting the chromosomes of *Rosa* 'Old Blush' to highlight the contribution of its putative parental lineages.

2.10.2 Materials and methods

We used the new version of the locus assembler aTRAM (Allen et al., 2018) to target the assembly of the 1856 SCO$_{\text{Tag}}$s in five unassembled diploid genomes of putative parents of *R.* 'Old Blush': *R. chinensis* var. *mutabilis* (GenBank ID: SRX3286282), *R. chinensis* var. *sanguinea* (GenBank ID: SRX3286285), *R. chinensis* var. *spontanea* (GenBank ID: SRX3286289), *R. gigantea* (syn. *R.* × *odorata* var. *gigantea*) (GenBank IDs: SRX3286283-SRX3286284) and *R.* × *odorata* 'Hume's Blush' (syn. *R. odorata* var *odorata*) (GenBank ID: SRX3286293), all published in Raymond et al. (2018). These five specimens correspond to ancient varieties that were in cultivation for centuries, first in China and then in Europe. It is thus difficult to be certain that the botanic varieties (*R. chinensis* var. *spontanea* or *R. odorata* var. *gigantea*) correspond to wild type specimens. Even for ancient varieties (*R. chinensis* var. *mutabilis*, *R. chinensis* var. *sanguinea*, *R.* × *odorata* 'Hume's

Blush'), their correct identification is not guaranteed since there were extensively exchanged and possibly crossed for breeding. Nevertheless, we took these accessions as they were and performed our analyses considering that their names truly reflect their genotypes. If different alleles were present at a $\mathrm{SCO_{Tag}}$ locus within one accession, only the first best blast hit was kept for further comparisons.

We also searched the heterozygous genome sequence of *R.* 'Old Blush' (Raymond et al., 2018), here after referred as OBhet, for the presence of the 1856 $\mathrm{SCO_{Tag}}$s using a reciprocal best blast analysis. Some $\mathrm{SCO_{Tag}}$s may present allele sequences in OBhet that could highlight the parental origins of *R.* 'Old Blush' when compared to the sequences recovered in the five putative parental Chinenses lineages. Therefore, only $\mathrm{SCO_{Tag}}$s with two different allele sequences in OBhet were kept for the study. To avoid bias associated with missing data, we discarded $\mathrm{SCO_{Tag}}$s that did not present both five parental sequences (one for each of the five putative progenitors) and two OBhet allele sequences. For each $\mathrm{SCO_{Tag}}$, the OBhet allele sequences were pairwise-compared to the five parental sequences using pairwise-distance (p-distance). The p-distance between two sequences was calculated as follow: p-distance = 1-S, with:

$$S = \frac{M}{N_{pos} + Gap \times \gamma}$$

M: Number of matches between the two sequences; $\mathrm{N_{pos}}$: Number of positions in the alignment; Gap: number of gap windows in the alignment; γ: gap penalty, set to -1.

Pairwise distances were then used to compute (1) the proximity between OBhet alleles and the five parental lineages, and (2) to perform a Principal Component Analysis (PCA) to identify possible grouping patterns. The proximity of each parental variety was calculated as the number of time that the corresponding parental lineage was found closest to OBhet alleles across the retained $\mathrm{SCO_{Tag}}$s. In case two or more parental lineages were closest to a OBhet allele, each received a fraction of the count that correspond to $1/x$, with x the number of same scoring parental lineages. The PCA was performed in R, using the package FactoMineR (Lê et al., 2008) and the function *PCA*. The groups were identified using the hierarchical clustering on principal components approach implemented in the function *HCPC*{FactoMineR}. Results were plot with the R package factoextra (Alboukadel and Mundt, 2017).

To further investigate the genomic contribution of the five parental lineages, we compared the 1856 $\mathrm{SCO_{Tag}}$ sequences of the haploid *R.* 'Old Blush' (OBhap) (Hibrand Saint-Oyant et al., 2018) with those found in the five Chinenses lineages using the p-distance as indicated above. The idea was to paint the seven pseudomolecules (idiogram) of OBhap with the contribution of each *Chinenses* lineages. Only complete $\mathrm{SCO_{Tag}}$ sets with five *Chinenses* alleles and one OBhap sequence were used. In case two or more parental lineages are closest to the sequence of OBhap, we discarded the $\mathrm{SCO_{Tag}}$ for chromosome painting. We considered that the genomic region after each $\mathrm{SCO_{Tag}}$ and before the next one was colored with the corresponding parental $\mathrm{SCO_{Tag}}$ allele.

Table 3: Results from the assembly of the 1856 single-copy orthologous tags in five *Chinenses* varieties. Not found: Number of $\mathrm{SCO_{Tag}}$s that were not found in full length, proportions are relative to the total number of blasted $\mathrm{SCO_{Tag}}$s (1856). $\mathrm{SCO_{Tag}}$ allele proportions are relative to the sum of $\mathrm{SCO_{Tag}}$s that were subjected to a RBB. Total >1 allele: Number of $\mathrm{SCO_{Tag}}$s with at least one allele, the proportion is relative to the total number of $\mathrm{SCO_{Tag}}$ (1856). RBB: Reciprocal Best BLAST.

Species	Not found	# $\mathrm{SCO_{Tag}}$s allele recovered after RBB search					Total > 1 allele
		0	1	2	3	>3	
R. chinensis var. *mutabilis*	35 (2%)	23 (1%)	1455 (80%)	307 (17%)	31 (2%)	5 (0%)	1798 (97%)
R. chinensis var. *sanguinea*	9 (0%)	9 (0%)	1639 (89%)	179 (10%)	18 (1%)	2 (0%)	1838 (99%)
R. chinensis var. *spontanea*	15 (1%)	14 (1%)	1750 (95%)	68 (4%)	9 (0%)	0 (0%)	1827 (98%)
R. gigantea	37 (2%)	16 (1%)	1656 (91%)	136 (7%)	10 (1%)	1 (0%)	1803 (97%)
R. × odorata	6 (0%)	14 (1%)	1470 (79%)	315 (17%)	48 (3%)	3 (0%)	1836 (99%)

2.10.3 Results

$\mathrm{SCO_{Tag}}$s assembly

The target assembly of the 1856 $\mathrm{SCO_{Tag}}$s was very efficient and enabled to recover at least one allele at more than 97% of the set of $\mathrm{SCO_{Tag}}$s across the five putative parental lineages (Table 3; Additional file S1). Most of the time, only one sequence was found per $\mathrm{SCO_{Tag}}$ locus, although the number of single sequence $\mathrm{SCO_{Tag}}$s varies from 1455 for *R. chinensis* var. *mutabilis* to 1750 for *R. chinensis* var. *spontanea*. For all five varieties, the average percentage of $\mathrm{SCO_{Tag}}$ heterozygosity (ie $\mathrm{SCO_{Tag}}$ with two different sequences) is 11% of the number of found $\mathrm{SCO_{Tag}}$s (from 4% in *R. chinensis* var. *spontanea* to 17% in *R. × odorata*). The BLAST search performed on OBhet genome sequences yielded 1762 $\mathrm{SCO_{Tag}}$s with two best hits found in different locations, of these 1278 $\mathrm{SCO_{Tag}}$s show different allele sequences, meaning that 72.5% of the $\mathrm{SCO_{Tag}}$ loci found with two sequences are heterozygous. Finally, 1174 $\mathrm{SCO_{Tag}}$s exhibit a complete set of sequences with five sequences from the parental lineages and two allele sequences from OBhet. Regarding OBhap, a total of 1702 $\mathrm{SCO_{Tag}}$s with five *Chinenses* alleles and one OBhap sequence were recovered for chromosome painting.

Parental lineages of R. 'Old Blush'

We used a measure of pairwise p-distance between OBhet alleles and their corresponding *Chinenses* alleles (Additional file S2) to calculate the proximity of the five putative parental lineages to OBhet alleles (Figure 30A). We observed that *R. × odorata* is the closest parental lineage to OBhet, with an overall proximity score of 671. The second closest parental lineage is *R. chinensis* var. *mutabilis* with a proximity score near 546. However, *R. chinensis* var. *sanguinea* displays a proximity score of 502, nearby that of *R. chinensis* var. *mutabilis*. *R. chinensis* var. *spontanea* has an intermediate proximity score of 373 and *R. gigantea* is the least close to OBhet with a proximity score of 256. The total proximity score of the *R. chinensis* varieties reaches 1421 (60% of the sum of proximities) while the total proximity score of the two *R. odorata* varieties reaches 927 (40% of the sum of proximities). The first two components of the PCA successfully captured 79% of the total data variance (Figure 30B). We observed three main variable types in the PCA plot (Figure 30B). *R. chinensis* var. *mutabilis* (CHIMU) and *R.* × *odorata* (ODO) form two distinct variables in the top right hand quarter while the variables linked to *R. chinensis* var. *spontanea* (CHISP), *R. chinensis* var. *sanguinea* (CHISA) and *R. gigantea* (GIG) are highly correlated in the bottom right hand quarter and could be merged in one variable. This conveys the idea that only *R. chinensis* var. *mutabilis* and *R.* × *odorata* influence greatly the data as observed in Figure 30A. The three highly correlated variables (CHISP, GIG and CHISA) may just reflect $\mathrm{SCO_{Tag}}$ alleles that are so conserved that they score similarly between those three varieties.

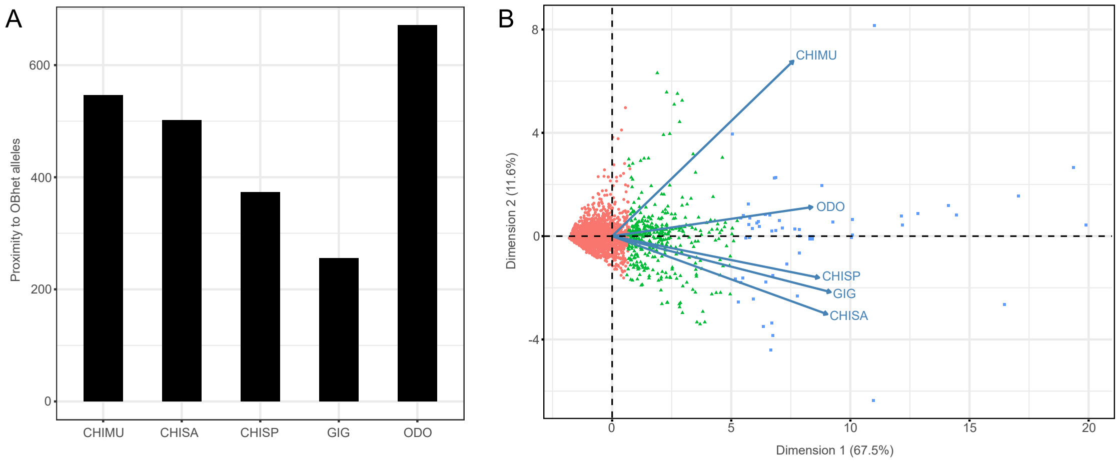


Figure 30: Close relatives of heterozygous *Rosa chinensis* 'Old Blush'. **A**. Bar plots of parental lineage proximities toward heterozygous $\mathrm{SCO_{Tag}}$s of *R.* 'Old Blush'. **B**. Principal Component Analysis of the p-distances between heterozygous alleles of *R.* 'Old Blush' and the alleles from the five *Chinenses* lineages. Each point represents an allele from the heterozygous *R.* 'Old Blush'.

Genomic contribution of the parental lineages in haploid *R.* 'Old Blush' idiogram

To better characterize the contribution of the five putative parental lineages of *R.* 'Old Blush', we calculated the proximity of each of the *Chinenses* alleles towards the OBhap allele identified in Debray et al. (2019). Only $\mathrm{SCO_{Tag}}$s with one minimum p-distance value were kept, thus 738 $\mathrm{SCO_{Tag}}$s contributed to the chromosome painting (Figure 31). We observed that the idiogram is quite fragmented, with portions inherited from the five putative parental lineages. In detail, some areas seem to be inherited only from specific lineages. For instance, the middle of LG 2, including the centromeric region, is exclusively from *R. chinensis* varieties. This is also the case for LG 5 and LG 7, that display large regions inherited from *R. chinensis* varieties. The contribution of *R.* × *odorata* is more dispersed. Some large portions inherited from *R.* × *odorata* may be found at the beginning of LG 1 and LG 3 and the second part of LG 6. It is worth mentioning that some areas are not well covered by $\mathrm{SCO_{Tag}}$s, and generally correspond to pericentromeric regions. This is especially true for the beginning of LG 3 that displays no $\mathrm{SCO_{Tag}}$s. This is explained by the way $\mathrm{SCO_{Tag}}$s are identified. The method only targeted variable regions in the Rosa genomes, so areas that are not well covered may correspond to highly conserved genomic portions across the genus *Rosa*.

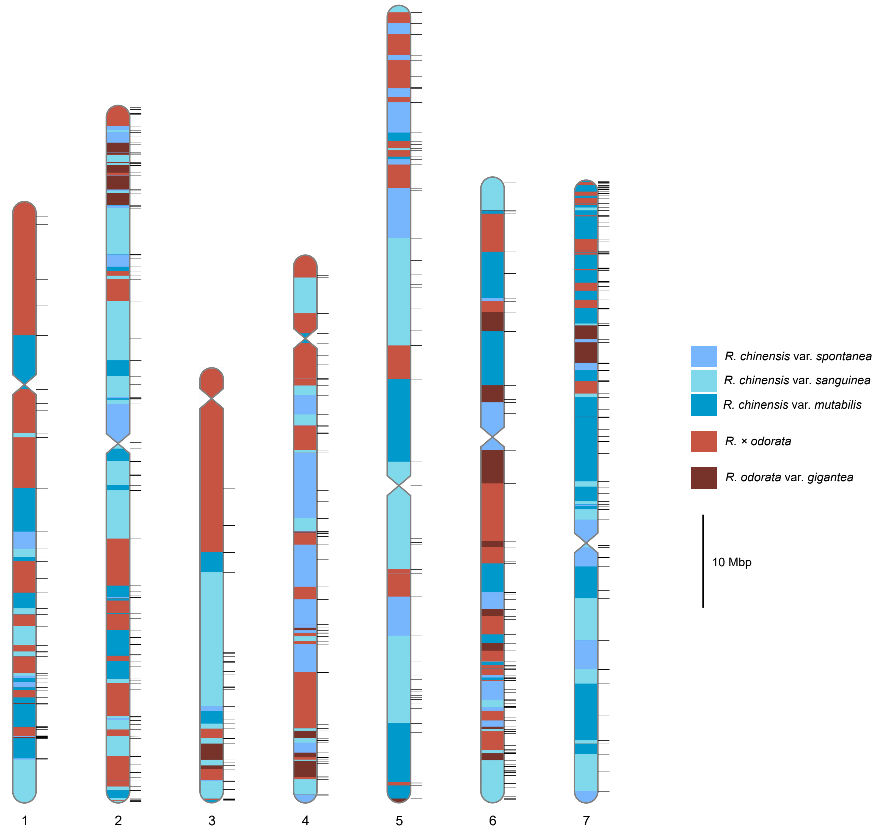

R. chinensis var. spontanea
R. chinensis var. sanguinea
R. chinensis var. mutabilis
R. × odorata
R. odorata var. gigantea
10 Mbp
1
2
3
4
5
6
7

Figure 31: Chromosome painting of haploid *Rosa chinensis* ‘Old Blush’. The seven Linkage Groups (LG) are colored according to $\mathrm{SCO_{Tag}}$s least proximity with one of the five *Chinenses* lineages. Each tick represents the position of at least one $\mathrm{SCO_{Tag}}$. One tick can mask the presence of multiple consecutive $\mathrm{SCO_{Tag}}$s. The pericentromeric regions are indicated by narrow areas on the idiograms, following Hibrand Saint-Oyant et al. (2018).

2.10.4 Discussion

$\mathrm{SCO_{Tag}}$s sequence variations were useful to study the origin of *R.* ‘Old Blush’. Using the updated version of aTRAM (Allen et al., 2018), $\mathrm{SCO_{Tag}}$s were easily assembled from Whole Genome Sequence (WGS) read data. This enabled to handle a large set of $\mathrm{SCO_{Tag}}$s with no missing data and perform a thorough comparison of allele variations between *R.* ‘Old Blush’ and its five putative parental lineages of section *Chinenses*.

The origin of *R.* ‘Old Blush’ is not straightforward. We found that *R.* ‘Old Blush’ results from a mix between *R. chinensis*-like and *R. odorata*-like lineages, in line with previous observations (Meng et al., 2011). Considering that the three *R. chinensis* varieties form a group, the first contributor to *R.* 'Old Blush' may be an interbred specimen of *R. chinensis*. The second contributor may be a variety of *R. odorata*, namely *R. odorata* 'Hume's Blush' as referenced in Raymond et al. (2018). The accession of *R.* × *odorata* 'Hume's Blush' (*R. odorata* var. *odorata* in Meng et al. (2011)) that we used here may correspond to one of the first Chinese varieties introduced in Europe at the turn of the 18th century. However, correct naming of ancient rose varieties as well as their purity is difficult to assess after centuries of extensive trade and breeding. *R. gigantea* (possibly *R. odorata* var. *gigantea*) seems to have a minor contribution to *R.* 'Old Blush'. Among the *R. chinensis* lineages, *R. chinensis* var. *mutabilis* seems to have largely contributed to the nuclear genome of *R.* 'Old Blush', while *R. chinensis* var. *spontanea* is the least related to *R.* 'Old Blush'. Yet, *R. chinensis* var. *spontanea* is often found the closest parent to *R.* 'Old Blush' in plastid analyses. This conveys the idea that the maternal parent of *R.* 'Old Blush' may better correspond to an interbred specimen of *R. chinensis* varieties. Here, we could not conclude that *R.* 'Old Blush' is the result from a single cross between contrasted *R. chinensis* var. *spontanea* and *R. odorata* var. *gigantea* as it is often assumed. Instead, *R.* 'Old Blush' seems to result from interbred varieties of *R. chinensis* and *R. odorata*.

Our analysis relies on the assumption that all five *Chinenses* lineages could be considered as putative parents of *R.* 'Old Blush' although this assumption may be baseless. Indeed, in a recent study combing SSR genotyping and karyotyping in old Chinese garden roses, Tan et al. (2017) concluded that the Old Blush group is the most primitive one. They suggested that rose cultivars evolved from the Old Blush group (itself derived from crosses between wild lineages) to the *Odorata* group (including *R.* × *odorata* 'Hume's Blush'), the Ancient Hybrid China group (including *R. chinensis* var. *mutabilis* and possibly *R. chinensis* var. *sanguinea*) and the modern roses. Future studies focusing on cultivars' origins such as the one developed here will have to (1) consider the way of crossing and (2) include many more accessions per species and variety to strengthen their conclusions. For the last point, future studies should even consider to first characterize the pool of putative parental lineages through genetic diversity and structure analyses.

Our set of $\mathrm{SCO_{Tag}}$s enabled to paint the idiogram of haploid *R.* 'Old Blush' with the contributions of the respective putative parental lineages. The results again highlight the mixed origin of *R.* 'Old Blush', with many fragmented portions inherited from the *Chinenses* lineages. Some areas were not well covered with $\mathrm{SCO_{Tag}}$s and it would probably be interesting to also find genomic information in these regions. $\mathrm{SCO_{Tag}}$s proved to be useful for chromosome painting in section *Chinenses*, although they were developed for the broad genus *Rosa*. We therefore think that $\mathrm{SCO_{Tag}}$s would be even more useful to study the genomic contribution of parental lineages in the case of an intersectional hybridization. However, the main disadvantage of $\mathrm{SCO_{Tag}}$s is that their positions were inferred based on one reference genome, thus hindering a proper study of synteny and genome rearrangements with other species. Such rearrangements may be particularly important during interspecific hybridizations, with or without ploidy increase, through the jump and/or expansion of transposable elements over chromosomes (Ungerer et al., 2006; Bashir et al., 2018; Latta et al., 2019). However, only long read sequencing of additional *Rosa* species combined with the identification of a set of milestone $\mathrm{SCO}_{\mathrm{Tag}}$s would enable to properly compare genome breaks between species.

2.10.5 Conclusion

Bridging the gap between cultivated materials and their corresponding wild counterparts is of great interest to understand the origin of certain alleles. Thanks to their distribution throughout the *Rosa* genome, $\mathrm{SCO}_{\mathrm{Tag}}$s represent valuable and cost-effective alternatives to WGS and assemblies to study the contribution of (wild) lineages to a known hybrid. Here, we demonstrated that *R.* 'Old Blush' is a hybrid between several interbred varieties of *R. chinensis* and *R. odorata* and does not result from a simple cross between the two wild species *R. chinensis* var. *spontanea* and *R. odorata* var. *gigantea*. We advocate that future studies should (1) include more accessions per species and variety to strengthen their conclusions and (2) clarify the orientation of crosses through genealogy with possible cross checking with historical data in the case of cultivated hybrid. If associated with long read sequencing, we think that $\mathrm{SCO}_{\mathrm{Tag}}$s comparisons in a situation that involves two contrasted species and their intersectional hybrid would certainly shed new light on the genomic rearrangements following such genomic shock.

2.10.6 Additional file and availability of data and material

Additional file S1. All $\mathrm{SCO}_{\mathrm{Tag}}$s sequences from the five *Chinenses* lineages recovered using aTRAM (Fasta file).

Additional file S2 Sequences from the 1174 $\mathrm{SCO}_{\mathrm{Tag}}$s with complete taxon occupancy (Fasta file). Complete taxon occupancy means that each $\mathrm{SCO}_{\mathrm{Tag}}$ is represented by five *Chinenses* alleles and two alleles from the heterozygous genome of *R.* 'Old Blush'

The two aforementioned Additional files are available upon request to the GDO team.

3 Unveiling the patterns of reticulated evolutionary processes with phylogenomics: Hybridization and polyploidy in the genus *Rosa* as models

3.1 Preamble

In chapter 2, we developed a general method for identifying and assessing single-copy orthologous tags in the *Rosa* genomes. Such tags fulfill most of the conditions required for phylogenetic analyses. $\mathrm{SCO_{Tag}}$s correspond to independent 300-600 bp orthologous sequences that are well-distributed along the seven pseudomolecules of the reference genome. The phylogenetic informativeness of each $\mathrm{SCO_{Tag}}$ has been studied to assess their ability to resolve ancient and recent speciation events. Chapter 2 identified 1856 $\mathrm{SCO_{Tag}}$s across 17 *Rosa* species/varieties. In the present chapter, we selected a subset of 92 nuclear $\mathrm{SCO_{Tag}}$s among the 1856 $\mathrm{SCO_{Tag}}$s, that were able to cover the 30 MY of divergence of the genus *Rosa*. Targeting nuclear $\mathrm{SCO_{Tag}}$s enables to access allelic information that may be used to infer the parental origins of hybrid diploids and allopolyploids. In addition, we developed four informative $\mathrm{SCO_{Tag}}$s in the chloroplast genomes to recover a plastid phylogeny that would embrace all accessions of the study. In the present chapter, the selected $\mathrm{SCO_{Tag}}$s were targeted in a broader taxon sampling than in chapter 2, covering about 120 species and varieties in the genus *Rosa*. The aim of this chapter is to produce a robust phylogenetic hypothesis for the genus *Rosa*, considering both hybridization and polyploidy as major driving forces.

This work has not been published yet but is under the submission process. The first part of this chapter correspond to the main article. The authors who contributed to this work are:

Kevin Debray[1], Marie-Christine Le Paslier[2], Aurélie Bérard[2], Tatiana Thouroude[1], Gilles Michel[1], Jordan Marie-Magdelaine[1], Fabrice Foucher[1] and Valéry Malécot[1]

[1] IRHS, Agrocampus-Ouest, INRA, Université d'Angers, SFR 4207 QuaSaV, Beaucouzé, France

[2] Etude du Polymorphisme des Génomes Végétaux (EPGV), INRA, Université Paris-Saclay, 91000 Evry, France

The contribution of each author is detailed at the end of the first part of this chapter. The second part of this chapter deals with the chloroplast whole genome phylogeny of the genus *Rosa* along with updated proposals for future taxonomic revisions in this genus.

3.2 Introduction

Phylogenetic trees are the current and most common way to illustrate the evolutionary history of organisms. They provide a pattern that consists of dichotomies, although the evolutionary processes underlying the diversification in a group are not limited to tree-like patterns. Bifurcating trees may therefore not represent all the processes that create diversity. At the same time, phylogenetic trees are increasingly used as a means to revise classifications so that such classifications could reflect evolutionary proximity between members of the group. Various authors have discussed the way to represent the evolutionary history of a group, based mainly on genealogy (Sosef, 1997; Welzen, 1997; Brummitt, 2002; Hörandl, 2006; Aubert, 2015). Others have produced tools to reconstruct phylogenetic networks that are assumed to best represent reticulations (Huson and Bryant, 2006; Solís-Lemus et al., 2017; Wen et al., 2018). Nevertheless, it is known that phylogenetic trees represent the history of genes rather than the history of organisms (Doyle, 1992; Maddison and Wiens, 1997). With the advent of sequencing, access to molecular data has been greatly facilitated, making it possible to handle large datasets potentially based on full genome sequences. Building a comprehensive evolutionary history of organisms capable of considering both reticulation and bifurcation processes still represents a major challenge, especially for large taxonomic groups. This study aims at developing a global framework to reconstruct both diverging and reticulate processes in a wide and complex taxonomic group, using large arrays of molecular sequences. We chose the genus *Rosa* to develop our framework since it represents a challenging group for both evolutionary biologists and taxonomists. There are several inherent aspects of *Rosa* that hamper the development of a comprehensive evolutionary history for this genus: (1) Reproductive barriers are loose within the genus *Rosa* and interspecific hybridizations are very common (Joly and Bruneau, 2007; Herklotz and Ritz, 2017; Vaezi et al., 2019), along with odd modes of character inheritance (Nybom et al., 2004; Wissemann and Ritz, 2007). This results in a large and complex genus of 150-200 intertwined species (Rehder, 1940; Wissemann, 2003a) that are difficult to precisely delineate; (2) A wide range of ploidy levels exists in *Rosa*. About half of the species are diploid, whereas the remaining are polyploid with almost all odd and even ploidy levels from triploid (3x) to decaploid (10x) (Hurst, 1925; Roberts et al., 2009; Jian et al., 2010), and sometimes including multiploid species (Lewis, 1959). Some sections are assumed to encompass only polyploid species with mixed characteristics that hinder correct species assignment (Wissemann, 1999; Ritz et al., 2005); (3) Many people have focused on *Rosa* for centuries, leading to a tremendous number of descriptions and artificial hybrids that added even more confusion to the already twisted evolutionary history of *Rosa* (Brumme et al., 2013; Masure, 2013).

There have been numerous attempts to understand the evolutionary history of *Rosa* based on molecular data (Matsumoto et al., 1998; Jan et al., 1999; Wu et al., 2001; Wissemann and Ritz, 2005; Scariot et al., 2006; Bruneau et al., 2007; Koopman et al., 2008; Meng et al., 2011; Qiu et al., 2012, 2013; Fougère-Danezan et al., 2015). However, they generally used plastid sequences alone, although their unilateral inheritance limits the study of reticulations. When nuclear data were considered, they either represented phenetic relationships in the case of molecular markers or referred to only one sequence (*GAPDH* (Joly et al., 2006b; Fougère-Danezan et al., 2015), nrITS (Matsumoto et al., 2000a; Wu et al., 2001; Wissemann and Ritz, 2005), and were therefore biased toward gene tree evolution.

Here, we improve upon previous phylogenetic knowledge of *Rosa* by sequencing 96 informative loci across 142 individuals representing most *Rosa* species. This is one of the first studies that covers the reticulate aspects of the evolution on a large and complex dataset. We consider that the application of such a framework will help resolve other challenging genera or taxonomic groups (*Rubus*, *Gossypium*, *Pyrinae*), given that 25% of plant species are involved in interspecific hybridizations (Mallet, 2005) and 15% of angiosperm speciations are associated with a ploidy increase (Wood et al., 2009).

3.3 Materials and methods

3.3.1 Plant material and nomenclature

A total of 142 accessions representing 126 species and subspecies out of the 150-200 potential wild rose species were sampled for this study (Appendix B, Supplementary Figure B.1). We followed Rehder's classification (Rehder, 1940) that divides the genus *Rosa* into four subgenera (*R.* subgen. *Rosa*, *R.* subgen. *Hulthemia* (Dumort.) Focke, *R.* subgen. *Platyrhodon* (Hurst) Rehder and *R.* subgen. *Hesperhodos* Cockerell). Most wild roses belong to the subgenus *Rosa* that is further divided into ten sections (*R.* sect. *Pimpinellifoliae* (DC.) Ser., *R.* sect. *Gallicanae* (DC.) Ser., *R.* sect. *Caninae* (DC.) Ser., *R.* sect. *Carolinae* Crép., *R.* sect. *Rosa* [= *R.* sect. *Cinnamomeae* (DC.) Ser.], *R.* sect. *Synstylae* DC., *R.* sect. *Chinenses* Ser. [= *R.* sect. *Indicae* Thory], *R.* sect. *Banksianae* Lindl., *R.* sect. *Laevigatae* Thory and *R.* sect. *Bracteatae* Thory). In our study, we adopted the designation of *Rosa cinnamomea* L. (syn. *Rosa majalis* Herrm.) as the type species of the genus, a proposal of Jarvis (1992) and validated in 2005 at the Vienna International Botanical Congress. This implies that the section previously known as *Rosa* sect. *Cinnamomeae* (DC.) Ser. is renamed *R.* sect. *Rosa*, while the former autonymous section now carries the name *Rosa* sect. *Gallicanae*.

Fresh leaves were immediately dried after cutting using silica gel and stored at IRHS (Additional File 1). Accessions with a wild known origin were preferred over garden-grown specimens derived from seeds. To avoid the over-representation of some subgenera or sections, we respected the species proportions present in Rehder's classification as much as possible (Rehder, 1940). For species with a broad distribution or with a doubtful wild origin, we tried to sample several accessions from different places. Due to the tremendous number of synonyms in the *Rosa* nomenclature, we decided to adopt, whenever possible, the nomenclature proposed by Masure (2013) that relies on the studies done by Brumme et al. (2013).

To expand the taxon sampling for plastid phylogenies, we used sequences from extra *Rosa* accessions from previous sequencing projects (Appendix B, Supplementary Table B.1). When no complete chloroplast genome sequence was available, we first assembled genomic paired-end reads retrieved from GenBank into a plastid genome sequence using NOVOPlasty (Dierckxsens et al., 2017) with the plastid genome sequence of *R.* 'Old Blush' as the seed sequence. We then searched complete plastid genome sequences for the presence of our plastid markers using BLAST.

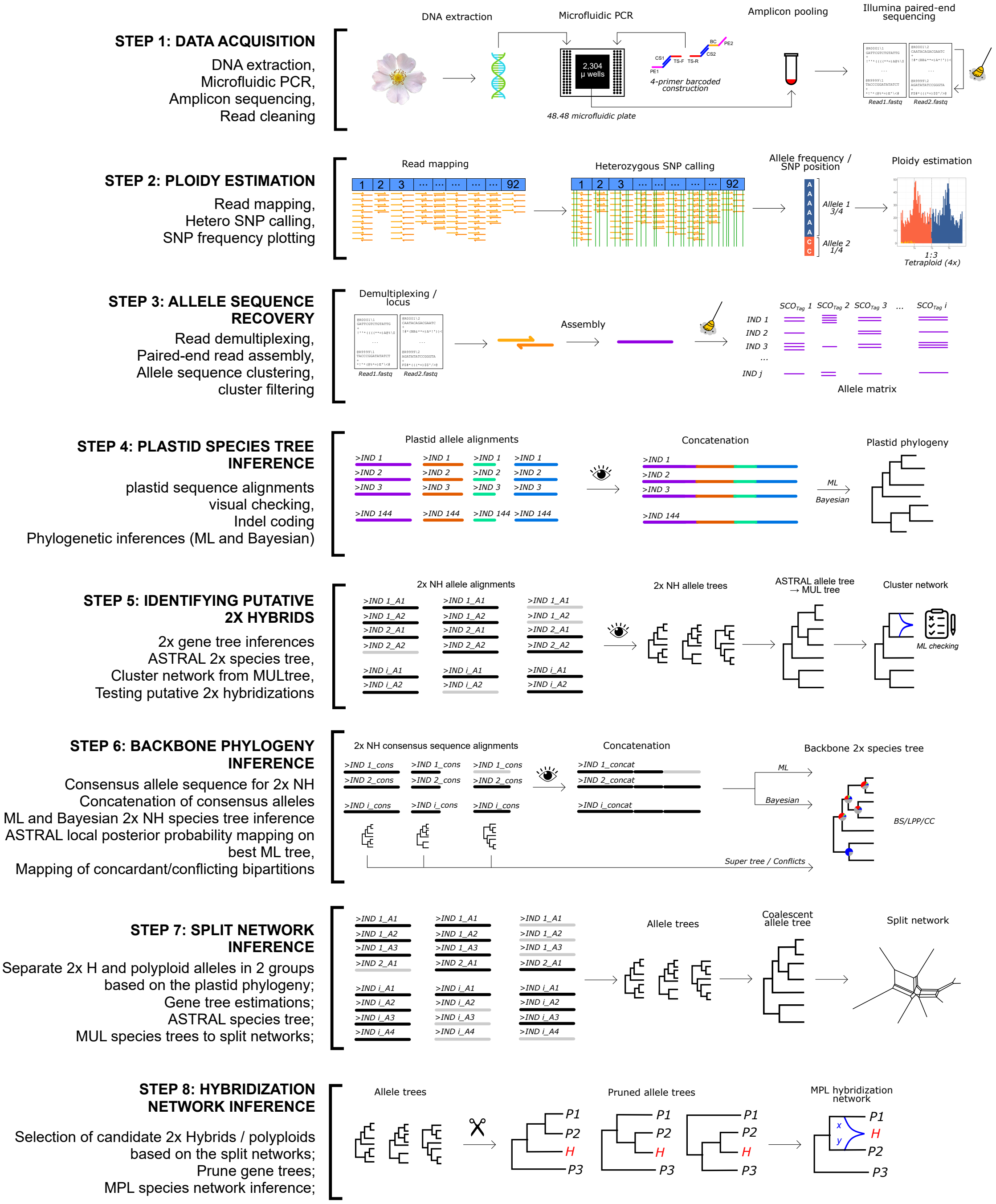
3.3.2 Selection of loci for phylogenomics

A total of 96 informative single-copy orthologous tags (SCO_{Tag}s) were chosen according to the method proposed in Debray et al. (2019) to resolve the phylogeny of *Rosa* as well as to study reticulate evolutions. We favored the selection of: (1) well-distributed SCO_{Tag}s along the seven chromosome sequences; (2) complementary SCO_{Tag}s regarding their phylogenetic informativeness along the 30 MY of divergence of the genus *Rosa*; and (3) SCO_{Tag}s that display an average level of concordance toward the species-tree topology proposed in Debray et al. (2019). Among the 96 selected SCO_{Tag}s, 92 were from the nuclear genome ($\text{nrSCO}_{\text{Tag}}$s from PAIR01 to PAIR92) and four from the plastid genome ($\text{cpSCO}_{\text{Tag}}$s from PAIR93 to PAIR96) (Additional File 2).

3.3.3 Extraction, amplification and sequencing

DNA was extracted using a modified version of the protocol proposed by Keb-Llanes et al. (2002). Modifications involved grinding 30 mg of dried tissues with a ball mill and tungsten beads, incubating ground tissues with extraction buffer for 1 h while mixing every ten minutes, setting the centrifuge to 13,200 rpm for the first round of centrifugation and to 10,200 rpm for the following, and resuspending the final pellet in 60 µL of TE buffer. Since plant tissues were collected from various *Rosa* species and leaf types at different times, the yield of DNA extraction may greatly vary from one sample to another. To standardize the extracted DNA samples for downstream analysis, we ran two quality checks. DNA purity and quantity were first assessed using a Nanodrop spectrophotometer. Extraction was considered to be satisfactory when both ratios of absorbance (A260/A280 and A260/A230) were above 1.8 and when DNA concentration was greater than 100 ng/µL. Since high DNA concentrations and good purity ratios may conceal fragmented DNA, samples that passed the previous test were further controlled with a PCR performed on a nuclear gene fragment of approximately 500 bp (*RoTFL1* gene, with primer pairs RoCen3-ACCACGAAGCCTAAGGTTGA and RoCen5-ATTTCACCACCTCCCTTCCT, as published in Remay et al. (2009)). A range of four dilutions (1/10, 1/25, 1/50 and 1/100) was used to dilute putative PCR inhibitors in extracted DNA and to find a dilution ratio able to amplify the control fragment for each DNA sample. Optimally diluted DNA samples were used for downstream amplification and sequencing experiments.

Since the SCO_{Tag} identification method involved the design of conserved PCR primer pairs, we used the microfluidic PCR amplification technique to amplify SCO_{Tag} alleles in the 144 DNA samples. Resulting amplicons performed by EPGV were then sequenced at great depth using


STEP 1: DATA ACQUISITION
DNA extraction,
Microfluidic PCR,
Amplicon sequencing,
Read cleaning
DNA extraction
Microfluidic PCR
2,304 µ wells
48.48 microfluidic plate
4-primer barcoded construction
Amplicon pooling
Illumina paired-end sequencing
Read1.fastq
Read2.fastq
STEP 2: PLOIDY ESTIMATION
Read mapping,
Hetero SNP calling,
SNP frequency plotting
Read mapping
Heterozygous SNP calling
Allele frequency / SNP position
Allele 1 3/4
Allele 2 1/4
Ploidy estimation
1:3
Tetraploid (4x)
STEP 3: ALLELE SEQUENCE RECOVERY
Read demultiplexing,
Paired-end read assembly,
Allele sequence clustering,
cluster filtering
Demultiplexing / locus
Assembly
Allele matrix
IND 1
IND 2
IND 3
IND j
STEP 4: PLASTID SPECIES TREE INFERENCE
plastid sequence alignments
visual checking,
Indel coding
Phylogenetic inferences (ML and Bayesian)
Plastid allele alignments
Concatenation
Plastid phylogeny
ML
Bayesian
STEP 5: IDENTIFYING PUTATIVE 2X HYBRIDS
2x gene tree inferences
ASTRAL 2x species tree,
Cluster network from MULtree,
Testing putative 2x hybridizations
2x NH allele alignments
2x NH allele trees
ASTRAL allele tree → MUL tree
Cluster network
ML checking
STEP 6: BACKBONE PHYLOGENY INFERENCE
Consensus allele sequence for 2x NH
Concatenation of consensus alleles
ML and Bayesian 2x NH species tree inference
ASTRAL local posterior probability mapping on best ML tree,
Mapping of concardant/conflicting bipartitions
2x NH consensus sequence alignments
Concatenation
Backbone 2x species tree
ML
Bayesian
BS/LPP/CC
Super tree / Conflicts
STEP 7: SPLIT NETWORK INFERENCE
Separate 2x H and polyploid alleles in 2 groups based on the plastid phylogeny;
Gene tree estimations;
ASTRAL species tree;
MUL species trees to split networks;
Allele trees
Coalescent allele tree
Split network
STEP 8: HYBRIDIZATION NETWORK INFERENCE
Selection of candidate 2x Hybrids / polyploids based on the split networks;
Prune gene trees;
MPL species network inference;
Allele trees
Pruned allele trees
MPL hybridization network
P1
P2
H
P3
x
y

(Caption next page.)

Figure 32: (Previous page.) Workflow of the analysis. **Step 1** represents the DNA extraction and the amplicon sequencing. PE1: Paired-end sequence 1; CS1: tag 1; CS2: tag 2; TS-F: Target-specific primer sequence forward; TS-R: Target-specific primer sequence reverse; BC: Barcode; PE2: Paired-end sequence 2. **Step 2** highlights the method used to infer ploidy level to each accession from the study of allele frequencies at nuclear heterozygous SNPs. This step involves the creation of a mapping reference made of the concatenation of single-copy orthologous tags (SCO_{Tag}s) extracted from the reference genome sequence of *R.* 'Old Blush'. **Step 3** illustrates the transition from partially demultiplexed reads to a matrix of filtered alleles. **Step 4** corresponds to the inference of plastid phylogenies using both Maximum Likelihood (ML) and Bayesian inferences. **Step 5** explains the procedure to identify non-hybrid diploid specimens (2x NH). Gray lines represent missing data. IND_1_A1 stands for Individual 1 allele 1. The multilabeled tree (MUL-tree), obtained after removing the allele info from leaf names in allele trees, is converted to a cluster network that reveals putative diploid hybridizations that are further tested in a ML framework. **Step 6** details the procedure to recover a robust backbone phylogeny of diploid putative progenitors. Input sequences are consensus alleles from non-hybrid diploids obtained from Step 5. Gray lines indicate missing data. ML and Bayesian analysis rely on a concatenation of all consensus allele sequences, whereas Astral coalescence and the Phyparts pipeline rely on individual allele trees. **Step 7** depicts the inference of split networks from trees containing consensus non-hybrid diploid alleles and alleles from hybrid diploids and polyploids. **Step 8** illustrates the inference of hybrid networks to test scenarios of hybridizations involving ten specimens or so (P_i) and a putative hybrid accession. (x, y) represent the contribution of each parental genome.

high-throughput sequencing at the Genoscope sequencing facility, Evry, France (step 1, Figure 32). The method proposed in Debray et al. (2019) for SCO_{Tag} identification already matches the requirements of the Fluidigm Access Array system for the design of primer pairs (uniform 60°C annealing temperatures, no homopolymers ≥ 3 bp). Two conserved sequences (CS1 and CS2, provided by the manufacturer) were added to the 5' end of both the forward and reverse primers to provide an annealing site for a second pair of primers corresponding to a concatenation of the complementary CS sequence, Illumina sequencing adapters, and a sample-specific barcode so that samples could be further multiplexed for the sequencing run. Microfluidic PCRs were performed in an Access Array System using six 48.48 Access Array integrated fluidic circuits, according to the manufacturer's protocol. Each array can simultaneously amplify 48 samples using each of 48 primer pairs individually. The resulting amplicons of the 2,304 PCRs were individually pooled before being sequenced on an Illumina MiSeq sequencer using 2 × 300-bp paired-end reads. We used six 48.48 Access Array integrated fluidic circuits to target the amplification of 96 SCO_{Tag}s across 144 individuals, rendering a total of 13,824 microfluidic PCRs. Two Illumina sequencing runs were performed to sequence the entire set of amplicons.

3.3.4 Read processing

Raw data paired-end reads were first processed by the Genoscope procedure to demultiplex read file output by the sequencer using the NGS barcode associated with each individual. This resulted in two fastq files (forward and reverse reads) for each individual where reads from all 96

$\mathrm{SCO_{Tag}}$s were mixed (end of step 1, Figure 32). We therefore ran a second read processing to further demultiplex individual reads by $\mathrm{SCO_{Tag}}$ and to trim reads of poor quality (step 3, Figure 32). Before demultiplexing, we made sure that all reads were correctly paired using custom script. We developed our own script to demultiplex reads by $\mathrm{SCO_{Tag}}$ (SCOtagsDemultiplexer.py). Each primer pair of length L is compared to the corresponding 5' fragment of length L of each read sequence, referred to as fore-read sequences. We computed the Levenshtein distance between the primer and the fore-read sequences and defined a threshold of 2, meaning that no more than two single-character edits (INDELs or substitutions) would be required to change one sequence into the other (–error option). In addition, we checked that the last four 3' bp of each fore-read sequence rigorously corresponded to the last four 3' bp of the primer sequence (–firmend option). Any read matching the above criteria was considered to be assigned to the corresponding primer. The process ended by a comparison between the demultiplexed forward and reverse read files to remove unpaired reads.

Demultiplexed reads were further trimmed, especially regarding their 3' ends, i.e., generally of lower quality and that could contain Illumina adapters if the read length exceeds the DNA insert size. We used Trimmomatic v0.32 (Bolger et al., 2014) for the quality trimming of reads by scanning the read with a 4-bp sliding window cutting when the average quality per base drops below 20. Illumina adapter sequences were retrieved from the fasta file TruSeq3-PE.fa and clipped using a seed mismatch of 2, a palindrome clip threshold of 30, and a simple clip threshold of 10. Reads with a length of less than 30 bp were discarded.

3.3.5 Ploidy estimations

The fact that tissue materials were silica-dried hinders the use of flow cytometry and/or chromosome count to accurately estimate the ploidy level of each individual. In addition, some samples were taken from previous studies where living tissue material was no longer available and would have led to uncertainty when referring only to the literature since multiploidy often occurs in *Rosa* and not all of the species have been described for their ploidy level. To circumvent these limits, we estimated ploidy levels based on the allele frequencies observed at heterozygous SNP positions (i.e., position with at least two alleles with a maximum frequency of 95% for the most present allele) after mapping all reads of one individual to a reference sequence (step 2, Figure 32). The mapping reference was made by concatenating all 92 selected nuclear $\mathrm{SCO_{Tag}}$ sequences retrieved from the reference genome sequence of *R.* 'Old Blush'. We modified the ploidyNGS.py script proposed by Augusto Corrêa dos Santos et al. (2017) in order to focus only on heterozygous SNP covered by at least ten reads. The distribution of heterozygous SNP frequencies was visually scored for each individual and compared to theoretical distributions to estimate the ploidy level. In the event of hesitation between several ploidy levels, the largest one was retained to represent the ploidy level of the accession for downstream analysis.

3.3.6 Allele recovery at each locus

Each forward read and its complementary reverse mate were assembled using FLASH2 Magoč and Salzberg (2011) (–min-overlap=10, –max-overlap=600, –mismatchRatio=0.25) to compute the corresponding allele sequence (step 3, Figure 32). Any read pair that could not be assembled due to missing overlap was artificially merged with eight N's. The resulting allele files were processed to match the input requirements of step 2 from the Pipeline for Untangling Reticulate Complexes (PURC) Rothfels et al. (2017) (see step 2 from the Fluidigm2PURC pipeline Blischak et al. (2018)) that aims to cluster allele sequences and remove chimera. This step provides a reduced number of putative allele sequences that are grouped into clusters. The original number of alleles belonging to each cluster determines its size. We performed three rounds of clustering and chimera detection using the modified version of the purc_recluster.py (purc_recluster2.py) proposed in Blischak et al. (2018). The similarity criteria were set to 0.995 and 0.997 for the first and second iteration, respectively. The minimum number of sequences per cluster for the cluster to be retained was set to [1, 5]. We used step 3 (crunch_clusters) from the Fluidigm2PURC pipeline (Blischak et al., 2018) to ultimately filter PURC clusters in a maximum likelihood framework, taking (1) the estimated ploidy level of each individual, and (2) the average per $\mathrm{SCO}_{\mathrm{Tag}}$ level of sequencing errors for all reads from that $\mathrm{SCO}_{\mathrm{Tag}}$ into consideration. Allele sequences were first realigned using mafft (Katoh and Standley, 2013) prior to crunching clusters (–realign flag). PURC clusters were then ranked from largest to smallest and the pipeline implemented a likelihood model for each cluster, with the cluster size as a variable and the ploidy level and the sequencing error per $\mathrm{SCO}_{\mathrm{Tag}}$ as parameters. Only the first K largest clusters were considered, K being the estimated ploidy level. The model associated with the maximum likelihood was selected and the most common allele sequence of each retained cluster was kept for downstream analysis (see Blischak et al. (2018) for detailed examples). Alleles were arbitrarily named from 1 to k where k is the number of clusters retained in the best ML model. The first allele corresponds to the representative allele sequence of the largest cluster. For the four plastid $\mathrm{SCO}_{\mathrm{Tag}}$s, we also used the crunch_clusters script to recover plastid sequences, except that we specified that the data is haploid (–haploid flag). All allele sequences used for downstream analysis are available in Additional File 3.

3.3.7 Reconstruction of a plastid phylogeny

Plastid sequences from all accessions were used to reconstruct a bifurcating plastid phylogeny that would give a first idea of the phylogenetic relationships between *Rosa* individuals, without taking reticulate evolution into account (step 4, Figure 32). We expected that this step would help to spot accessions with an odd placement, putatively due to misidentification or the hybrid origin of the sample.

Sequences from each plastid $\mathrm{SCO}_{\mathrm{Tag}}$s were aligned using mafft and visually assessed to remove misaligned regions generally due to repetitive nucleotide sequences. The best substitution model for each plastid $\mathrm{SCO}_{\mathrm{Tag}}$ was searched using jModelTest2 (Darriba et al., 2012). Alignments were then concatenated and INDELs were coded using the simple INDEL coding (SIC) method with 2matrix (Salinas and Little, 2014). Each original alignment was split into two partitions, one for the nucleotide alignment and one for the INDEL coding. This resulted in a concatenated alignment with eight partitions. Ten maximum likelihood searches of GARLI v2.0.1 (Zwickl, 2006) were performed to recover the best Maximum Likelihood (ML) tree, specifying the best substitution model found for each nucleotide partition. In addition, 1000 bootstraps replicates were used to assess the support of each branch of the best ML tree. We also conducted the same tree search with the same partitions and the same corresponding substitution models in a Bayesian framework in order to confirm the supports found in the ML analysis. To do this, we used MrBayes v3.2.7 (Huelsenbeck and Ronquist, 2001) for 100 million generations using three runs of four MCMC chains, each with a sampling of parameters every 10,000 generations. The burn-in fraction was estimated as 25% after visualizing parameter convergence in Tracer 1.7 (Rambaut et al., 2018). We used outgroup sequences of Fragaria nipponica, Fragaria vesca, Potentilla parvifolia, Potaninia mongolica, Drymocalis glandulosa, Comarum salesovianum and Sanguisorba officinalis from Genbank to root the plastid phylogenies (Appendix B, Supplementary Table B.1).

3.3.8 Identifying putative diploid hybrids

Individuals estimated as diploid were used to detect patterns of hybridization that could be further removed to draw a backbone phylogeny of putative diploid progenitors (step 5, Figure 32). All allele sequences from diploid specimens were aligned and visually checked to remove misaligned regions generally due to sequence repetitions. Allele trees were inferred with PhyML v3 (Guindon et al., 2003) using the best substitution model found for each alignment with jModelTest2 and 100 bootstrap replicates. Branches with less than 10% of bootstrap support were collapsed into polytomies. Phasing alleles from different $\mathrm{SCO_{Tag}}$s could not be done since alleles from each $\mathrm{SCO_{Tag}}$ were targeted independently. This implies that we could not know if allele 1 at locus A comes from the same parental genome as allele 1 from locus B, since allele 1 at each locus was given an arbitrary name according to the size of its original cluster. Unphased allele sequences did not normally hamper the detection of diploid hybrids since we expected that alleles from a non-hybrid specimen would fall into the same cluster. We therefore considered allele trees as gene trees with leaf labels corresponding to 'genome', i.e., able to represent genome relationships in the context of polyploidy events. ASTRAL v5.3.1 (Zhang et al., 2018a) was used to coalesce allele trees into a super allele tree, taking incomplete lineage sorting into consideration. Allele assignment of each leaf label was removed to convert the super allele tree into a super multilabeled tree (MUL-tree), meaning that multiple leaf labels share the same name that corresponds to the accession name. We visualized the super MUL-tree tree in Dendroscope v3 (Huson and Scornavacca, 2012) and applied the 'Cluster network' function to visualize reticulations among diploid accessions. Putative diploid hybridizations were further tested in a ML framework using the CalGTProb (Yu et al., 2014) from the PhyloNet v3.7.1 package (Than et al., 2008). Diploid genome trees were first pruned to retain the hybrid (H) and two of its putative progenitors (P1 and P2). Three scenarios were built, two of which considered H as the sister to either P1 or P2, and one that considered H as a hybrid between P1 and P2. The function computes the likelihood of each scenario, and the scenario with the highest likelihood was considered to best fit the underlying set of allele trees. Accessions found as hybrids in both the cluster network and the ML tests were discarded for the inference of the backbone phylogeny.

3.3.9 Inference of a backbone phylogeny

Allele alignments were trimmed to retain only alleles from diploid non-hybrid specimens to draw a backbone phylogeny (step 6, Figure 32). When two allele sequences were present at a locus, we merged them into a consensus sequence. Any INDEL or substitution between the two allele sequences was randomly selected to be included in the consensus sequence. INDELs of length L above 1 bp were included by a block of length L. Consensus allele sequences were used to infer different phylogenetic hypotheses using ML, Bayesian and coalescent methods. A fraction of the conflicting gene tree was plotted at each node of the backbone phylogeny using the PhyParts pipeline (Smith et al., 2015). We expect that the use of different methods for phylogenetic reconstruction would provide a thorough view of the support at each branch of the backbone tree. Consensus allele sequence alignments were concatenated using 2matrix.pl, and INDELs were coded as Simple Indel Coding (SIC). The partitioned alignment was used for both ML and Bayesian inferences. For ML inference, we used RAxML v.8 (Stamatakis, 2014) with a GTR substitution model with four gamma rates (+G) applied to each nucleotide partition, and a binary model (BIN) applied to each INDEL partition. We conducted 30 searches for the best ML tree and implemented 1000 bootstrap replicates to assess the support of each branch of the best ML tree. For Bayesian inference, we used three runs using four MCMC chains each in MrBayes. We ran the analysis for 100 million generations and a sampling of parameters every 10,000 generations. The burn-in fraction was estimated at 25% after visualizing the convergence of parameters in Tracer. The obtained topology was the same as the ML phylogeny, so we mapped clade credibility values to the ML backbone phylogeny. For coalescent-based analysis, each alignment of consensus diploid allele sequences served to estimate the corresponding best ML gene tree in PhyML using the best substitution model found in jModelTest2 and 100 bootstrap replicates. ASTRAL v5.3.1. was used to evaluate the support of each quadripartition (the four clusters around a branch) through local posterior probabilities (LPP). LPP varies between 0 and 1, and is based on the percentage of quartets in gene trees that agree or disagree with a branch of the species tree. In our case, we provided ASTRAL with the ML backbone topology to score each quadripartition. Finally, the fractions of concordant and conflicting gene tree topologies were plotted at each bipartition (node) of the backbone ML tree, using a BS threshold of 50% (i.e., gene tree bipartitions with less than 50% support were not considered for fraction estimations), so as to give an overall perspective of the underlying conflicts within the set of gene trees.

3.3.10 Inference of global split networks

Split networks were chosen to globally see the placement of diploid hybrids and polyploids compared to their putative diploid progenitors (step 7, Figure 32). To circumvent the computational burden associated with the analysis of a large dataset, we divided the genus into two parts: Meta Clade 1 (MC1) and Meta Clade 2 (MC2), based on the plastid phylogeny. Each meta clade is further subdivided into well-supported subclades/subgrades (SC). To ease phylogenetic computations and network visualizations, we drew two different split networks for hybrid diploids and polyploids of MC1 and MC2, respectively. Both analyses share a common base sampling that corresponds to all non-hybrid diploids and eight representative polyploids that cover each SC of the plastid phylogeny. Remaining hybrid diploids and polyploids are assigned to either MC1 or MC2 analysis based on their position in the plastid phylogeny. For each analysis, consensus non-hybrid diploid sequences, hybrid diploid and polyploid alleles were aligned to estimate the best ML tree with PhyML using the best substitution model found with jModelTest2 and 100 bootstrap replicates. Gene tree branches with less than 10% of bootstrap support were collapsed into polytomies before coalescent super tree estimation with ASTRAL. Leaf labels of the super coalescent tree were truncated to remove the allele numbering and obtain a MUL-tree. The MUL-tree was uploaded in Splitstree v4 (Huson and Bryant, 2006) and the ‘Consensus network’ algorithm was used to infer the split network.

3.3.11 Inference of hybridization networks

Hybridization networks aim at focusing on specific hybrid hypotheses that split networks may have arisen and provide much detail about the respective contribution of each parental lineage toward the hybrid specimen (step 8, Figure 1). We selected 12 putative hybrids, either diploid or polyploid, to further detail reticulate relationships between accessions. Accessions were selected based upon hybridization hypotheses made in previous studies and odd placement of the accession in either the plastid phylogeny or the split networks. We used the implementation of Maximum Pseudo Likelihood network inference (Yu and Nakhleh, 2015) in PhyloNet to estimate the best MPL network for each of the 12 accessions. Implementing the network with all alleles from all accessions was computationally intractable, so we selected around ten accessions that were tested as the putative parental lineages of the hybrid specimen based on: (1) previous studies; (2) the plastid phylogeny; and (3) the split networks. Allele trees were pruned to keep only alleles of putative parental accessions as well as alleles from the hybrid specimen. We assumed one hybridization event and specified the putative hybrid specimen. Alleles were mapped to their corresponding accessions. Branch lengths and inheritance probabilities were optimized during the MPL searches. We performed five MPL searches to recover the best MPL network with inheritance probabilities. We developed a custom script (HybridMapper.py) to prepare the nexus file input required for PhyloNet.

3.4 Results

3.4.1 The amplicon sequencing technique yields valuable allelic data for phylogenomics

The amplicon sequencing technique was used to recover allelic sequences in 144 accessions (Additional File 1) across 96 $\mathrm{SCO}_{\mathrm{Tag}}$s (Additional File 2) using two runs of Illumina sequencing (step 1, Figure 32). The sequencing yielded an average of 185k raw reads per accession, all $\mathrm{SCO}_{\mathrm{Tag}}$s included. This dropped by 20% to 149k reads per accession after read demultiplexing and trimming (step 3, Figure 32; Appendix B, Supplementary Figure B.2).

The amplicon sequencing approach successfully recovered a total of 26,894 alleles for phylogenomic analyses after filtering (Additional file 3). Eighty-eight percent of the microfluidic PCRs led to the assembly of at least one allele that could be further used in phylogenetic analyses (Appendix B, Supplementary Figure B.3). Seven accessions were removed for downstream analyses on nuclear markers due to either missing data in more than 50% of the nuclear $\mathrm{SCO}_{\mathrm{Tag}}$s (*R.* × *alba* (ALX01) and *R. yainacensis* (YAI01)) or disproportionate missing data for at least one of the two sequencing runs (*R. ecae* (ECA01), *R. henryi* (HEN02), *R. pimpinellifolia* (PIM01), *R. roxburghii* (ROX04) and *R. setipoda* (SEP03)) (Appendix B, Supplementary Figure B.4). The average taxon occupancy per $\mathrm{SCO}_{\mathrm{Tag}}$ was 88% (between 8% for PAIR_43 and 98% for PAIR_07). Five $\mathrm{SCO}_{\mathrm{Tag}}$s had taxon occupancy below 50% but were still conserved for downstream analysis.

3.4.2 Ploidy level estimations distinguish between diploid and polyploid accessions

Many ploidy levels exist in *Rosa* both at intra and interspecific levels, and knowing this information prior to phylogenetic analyses would ease the selection of diploid putative progenitors to reconstruct a backbone phylogeny for *Rosa*. Silica-dried materials like those used in this study hamper the application of wet-lab techniques for estimating ploidy levels (flow cytometry, chromosome count). To infer a ploidy level to each accession, we therefore used allele frequency distributions as a proxy for allele segregations (step 2, Figure 32). We mapped pre-processed reads of each accession on an artificial reference made of the concatenation of the 92 nuclear $\mathrm{SCO}_{\mathrm{Tag}}$ sequences extracted from the haploid reference genome sequence of *R.* 'Old Blush' (Hibrand Saint-Oyant et al., 2018). We analyzed allele frequencies at heterozygous SNPs covered by at least ten reads. Allele frequency distributions ease the recognition of diploid vs. polyploid accessions (Figure 33). Visual assignments to either of the two states resulted in scoring 75 accessions as diploid and 62 as polyploids (Appendix B, Supplementary Figure B.5). Accurate assignment of ploidy levels among polyploid accessions was more difficult and we therefore preferred to overestimate ploidy levels for accessions with a dubious allele frequency distribution in order to maximize the chance to recover meaningful allele sequences during the next steps of the analysis.

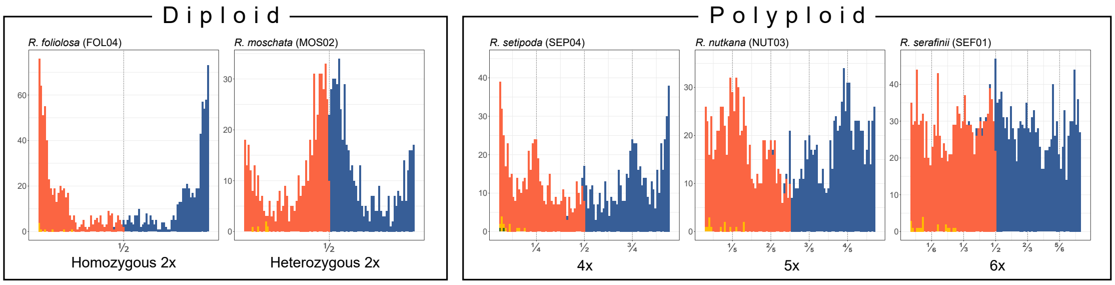


Figure 33: Ploidy assignation based on the distribution of allele frequencies at heterozygous SNP positions. Species names are followed by their accession code. Frequencies of the first, second, third and fourth allele at each heterozygous SNP position are represented by blue, orange, yellow and green colors, respectively. Allele frequency distributions are plotted within the range of 5-95%. Frequency distribution is shown on five examples. All distributions are presented in Appendix B, Supplementary Figure B.5.

3.4.3 Plastid sequences give a first and robust insight into phylogenetic relationships between *Rosa* specimens

We then wanted to have a first hypothesis regarding their phylogenetic relationships (step 4, Figure 32). We therefore used the four plastid SCO$_{\text{Tag}}$s to build a phylogenetic hypothesis for all accessions. We used ML and Bayesian reconstruction to compare branch supports and to obtain an overall idea of branch robustness. We observed that most of the deep branches are well supported in both analyses or at least in one of the two. The genus *Rosa* splits into two well-supported clades that we called Meta Clade 1 (MC1) and Meta Clade 2 (MC 2) in the ML analysis (Figure 34). However, the MC1 clade is not resolved in the Bayesian analysis and results in a polytomy at the base of the genus (C1+C2abc — C3 — MC2) (Appendix B, Supplementary Figure B.6). Most of Redher's sections are not monophyletic, with accessions assumed to belong to the same section that fall into distinct clades.

MC1 corresponds to a group that approximately encompasses *R.* sect. *Pimpinellifoliae*, *Rosa*, *Carolinae*, and, *Hesperhodos*, *Hulthemia* and *R.* subg. *Platyrhodon* pro parte. The *Pimpinellifoliae* section is not monophyletic. A main grade bears an Asian clade with bright-yellow flowered and flat prickled accessions (C1) and *R. foetida* that occupies an intermediate position between the main *Pimpinellifoliae* clade of C1 and a clade dominated by *R.* sect. *Rosa* (C2abc). All accessions of *R. spinosissima* (SPI05) and *R. pimpinellifolia* (PIM01, PIM03) fall into the same clade C2b, near accessions from *R.* sect. *Rosa*. Remaining *Pimpinellifoliae* accessions are found in the clades C2abc (*R. tsinglingensis* (TSI01), *R. kokanica* (KOK01) and *R. farreri* (FAR01)) or right in MC2 (*R. hemisphaerica* var *rapinii* (HES01), *R. koreana* (KOR02)). Most *Rosa* accessions are found within the C2abc clade that also includes all *Carolinae* accessions, except for *R. rugosa* (RUG02), *R. giraldii* (GIR02) and *R. bella* (BEL02) that are grouped with MC2 accessions and *R. moyesii* (MOY03) and *R. hemsleyana* (HEM02) found in the main *Pimpinellifoliae* clade (C1). All

Carolinae accessions are embedded in clade C2c that encompasses North American and European accessions from the section *Rosa*. The subg. *Platyrhodon* that usually includes two species (*R. praelucens* and *R. roxburghii*) is not monophyletic with *R. praelucens* (PAE01) found in MC1 and *R. roxburghii* (ROX03) in MC2. The subg. *Hulthemia* and *Hesperhodos* are monophyletic and found at the base of MC1.

MC2 includes *R.* sect. *Laevigatae*, *Bracteatae*, *Banksianae*, *Synstylae*, *Chinenses*, *Caninae*, *Gallicanae* and *R.* subg. *Platyrhodon* pro parte. The four mono/bi-species sections/subgenus (*R.* sect. *Laevigatae*, *Bracteatae*, *Banksianae* and *R.* subg. *Platyrhodon*) are found at the base of MC2 in a grade called C4. The rest of MC2 corresponds to a super clade of *R.* sect. *Synstylae* (Asian *Synstylae* in C5 and European *Synstylae* in C6b), *Chinenses* (C5), *Caninae* (C6ab) and *Gallicanae* (C5 and C6b). Two *Synstylae* species are found in MC1 (*R. glomerata* (GLO01) and *R. abyssinica* (ABY01)). All accessions of *R.* 'Old Blush' (OLD00-03) fall into the same clade, which also includes *R.* × *odorata* (ODO00) but not the rest of the *Chinenses* accessions that are spread over Asian *Synstylae* accessions. Accessions of *R.* × *damascena* (DAM00 and DAM02) are closer to *R. moschata* (MOS02) and *R. brunonii* (BRU01) than to other *Gallicanae* accessions found in clade C6b.

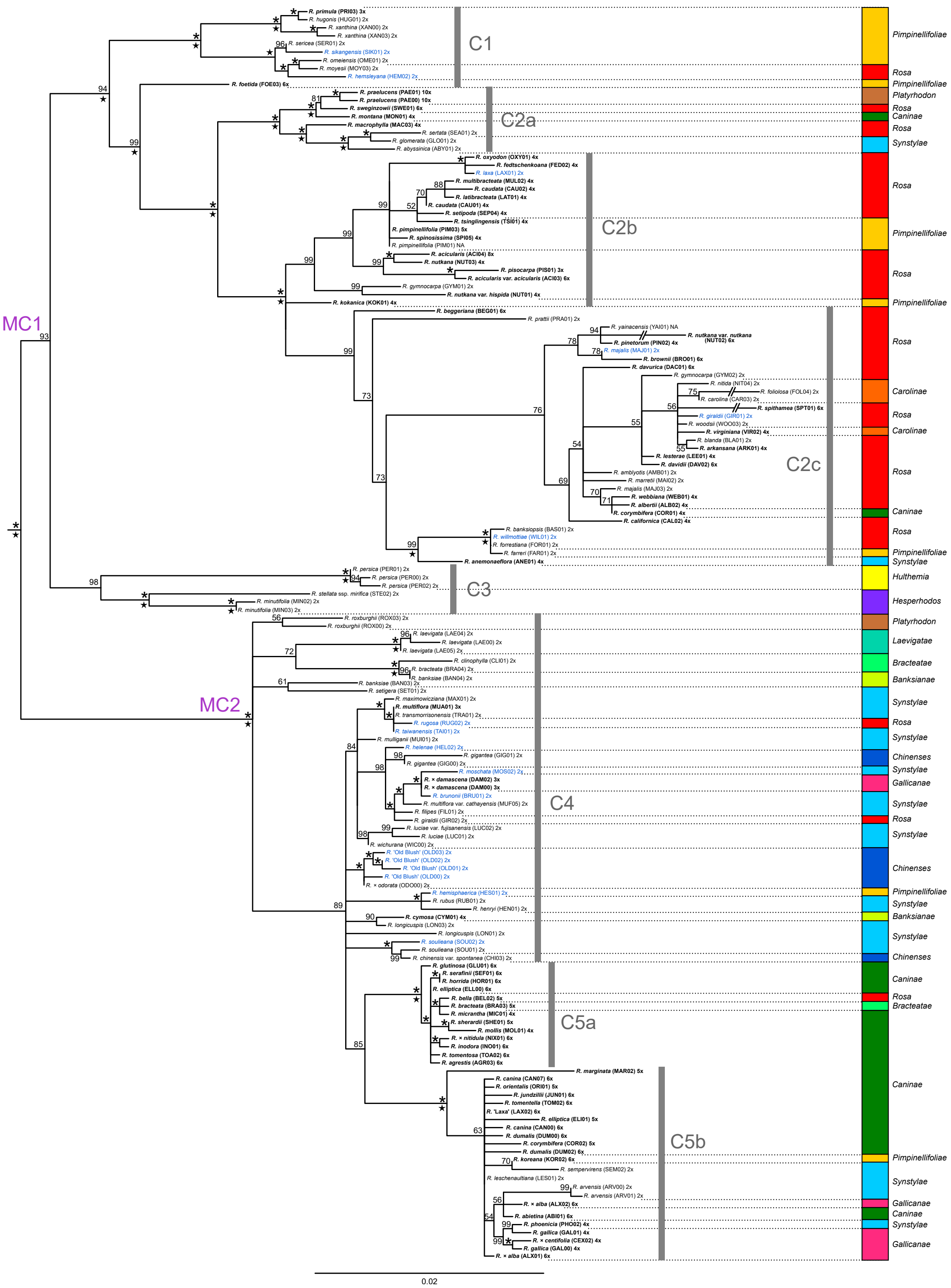
Accessions assumed to belong to a specific section/subgenus and that fall into distinct clades may correspond to hybrids, misidentified accessions or rootstock material if collected in rose gardens (see *R. bella* (BEL02), *R. bracteata* (BRA03), *R. cymosa* (CYM01), *R. hemisphaerica* (HES01), *R. koreana* (KOR02), *R. montana* (MON01), *R. rugosa* (RUG02)).

3.4.4 The backbone phylogeny highlights the origin of the diploid parental lineages

Unlike haploid plastid sequences, nuclear $\mathrm{SCO}_{\mathrm{Tag}}$s contain intra individual variability inherited by both parental lineages. Intra individual variations may reflect differences between subgenomes that were inherited from an earlier interspecific hybridization. Phylogenetic relationships are therefore better represented using networks than bifurcating trees when extensive interspecific hybridizations are suspected within a group. Prior to network inferences using nuclear data, we wanted to develop a backbone phylogeny for *Rosa* made of diploid accessions that would not have been identified as hybrids (steps 5 and 6, Figure 32). We assumed that conspecific nuclear alleles falling into the same clade represent an accession whose history does not involve any inter-lineage hybridization. We therefore infer allele trees and the corresponding diploid species tree to identify wandering alleles that may shed light on hybridization patterns. Incomplete lineage sorting (ILS) may also explain why conspecific alleles fall into different clades, so we searched for diploid hybrids in an ILS-aware frame to distinguish hybridization from ILS. We first inferred a consensus diploid network obtained from a transformation of the MUL-tree (itself derived from the Astral super allele tree). Sixteen putative diploid hybrids were found (Appendix B, Supplementary Figure B.7), resulting in 24 scenarios of hybridization involving one hybrid and two parents. Each scenario was individually tested in a ML ILS-aware frame. Nineteen hybridization scenarios were confirmed but

Table 4: Likelihood scores for diploid hybrid scenarios. # indicates the number of nuclear allele trees pruned to contain the alleles of Parent 1 (P1), Parent 2 (P2), the hybrid (H) and an outgroup species. S1 is the hybrid scenario, S2 and S3 are the null hypotheses where H is closer to P1 or P2, respectively. The maximum likelihood scores are in bold.

Hybrid	Parent 1	Parent 2	#	S1	S2	S3
R. sikangensis (SIK01)	*R. moyesii* (MOY03)	*R. hemsleyana* (HEM02)	83	-379.59	**-379.59**	-380.71
R. hemsleyana (HEM02)	*R. sericea* (SER01)	*R. sikangensis* (SIK01)	83	**-329.75**	-330.01	-329.89
R. majalis (MAJ01)	*R. woodsii* (WOO03)	*R. giraldii* (GIR01)	83	**-309.2**	-309.42	-309.20
R. giraldii (GIR01)	*R. blanda* (BLA01)	*R. majalis* (MAJ01)	53	**-193.13**	-193.62	-193.67
R. palustris (PAL04)	*R. carolina* (CAR03)	*R. foliolosa* (FOL04)	41	**-100.64**	-101.13	-108.98
R. laxa (LAX01)	*R. prattii* (PRA01)	*R. willmottiae* (WIL01)	46	-186.87	**-186.86**	-186.97
R. willmottiae (WIL01)	*R. laxa* (LAX01)	*R. forrestiana* (FOR01)	45	**-171.42**	-179.64	-180.24
R. willmottiae (WIL01)	*R. laxa* (LAX01)	*R. farreri* (FAR01)	46	**-190.16**	-192.12	-191.74
R. taiwanensis (TAI01)	*R. rubus* (RUB01)	*R. henryi* (HEN01)	60	**-231.17**	-235.85	-233.06
R. rugosa (RUG02)	*R. transmorrisonensis* (TRA01)	*R. luciae* (LUC01)	85	**-331.45**	-339.20	-335.02
R. rugosa (RUG02)	*R. transmorrisonensis* (TRA01)	*R. luciae* (LUC02)	84	**-342.58**	-346.52	-343.89
R. moschata (MOS02)	*R. brunonii* (BRU01)	*R. luciae* (LUC01)	81	**-283.65**	-291.14	-302.29
R. moschata (MOS02)	*R. brunonii* (BRU01)	*R. luciae* (LUC02)	81	**-296.89**	-303.25	-315.73
R. moschata (MOS02)	*R. brunonii* (BRU01)	*R. rugosa* (RUG02)	81	**-327.76**	-340.90	-344.90
R. moschata (MOS02)	*R. brunonii* (BRU01)	*R. transmorrisonensis* (TRA01)	79	**-288.67**	-297.67	-304.87
R. moschata (MOS02)	*R. brunonii* (BRU01)	*R. maximowicziana* (MAX01)	81	**-273.91**	-277.67	-285.09
R. hemisphaerica var *rapinii* (HES01)	*R. filipes* (FIL01)	*R. giraldii* (GIR02)	75	-250.97	-251.22	**-250.96**
R. hemisphaerica var *rapinii* (HES01)	*R. filipes* (FIL01)	*R. helenae* (HEL02)	74	**-261.20**	-262.69	-261.28
R. hemisphaerica var *rapinii* (HES01)	*R. filipes* (FIL01)	*R. multiflora* (MUF05)	74	-269.11	-269.42	**-269.10**
R. brunonii (BRU01)	*R. moschata* (MOS02)	*R. soulieana* (SOU02)	79	-298.53	**-298.46**	-310.95
R. soulieana (SOU02)	*R. brunonii* (BRU01)	*R. chinensis* (CHI03)	75	**-220.00**	-230.68	-224.10
R. 'Old Blush' (OLD02)	*R.* 'Old Blush' (OLD03)	*R.* 'Old Blush' (OLD01)	83	-447.94	-449.19	**-447.78**
R. 'Old Blush' (OLD01)	*R. chinensis* (CHI03)	*R. gigantea* (GIG01)	77	**-285.21**	-286.05	-285.73
R. 'Old Blush' (OLD02)	*R. chinensis* (CHI03)	*R. gigantea* (GIG01)	81	**-306.73**	-307.83	-307.52
R. 'Old Blush' (OLD03)	*R. chinensis* (CHI03)	*R. gigantea* (GIG01)	81	**-304.78**	-306.55	-306.08

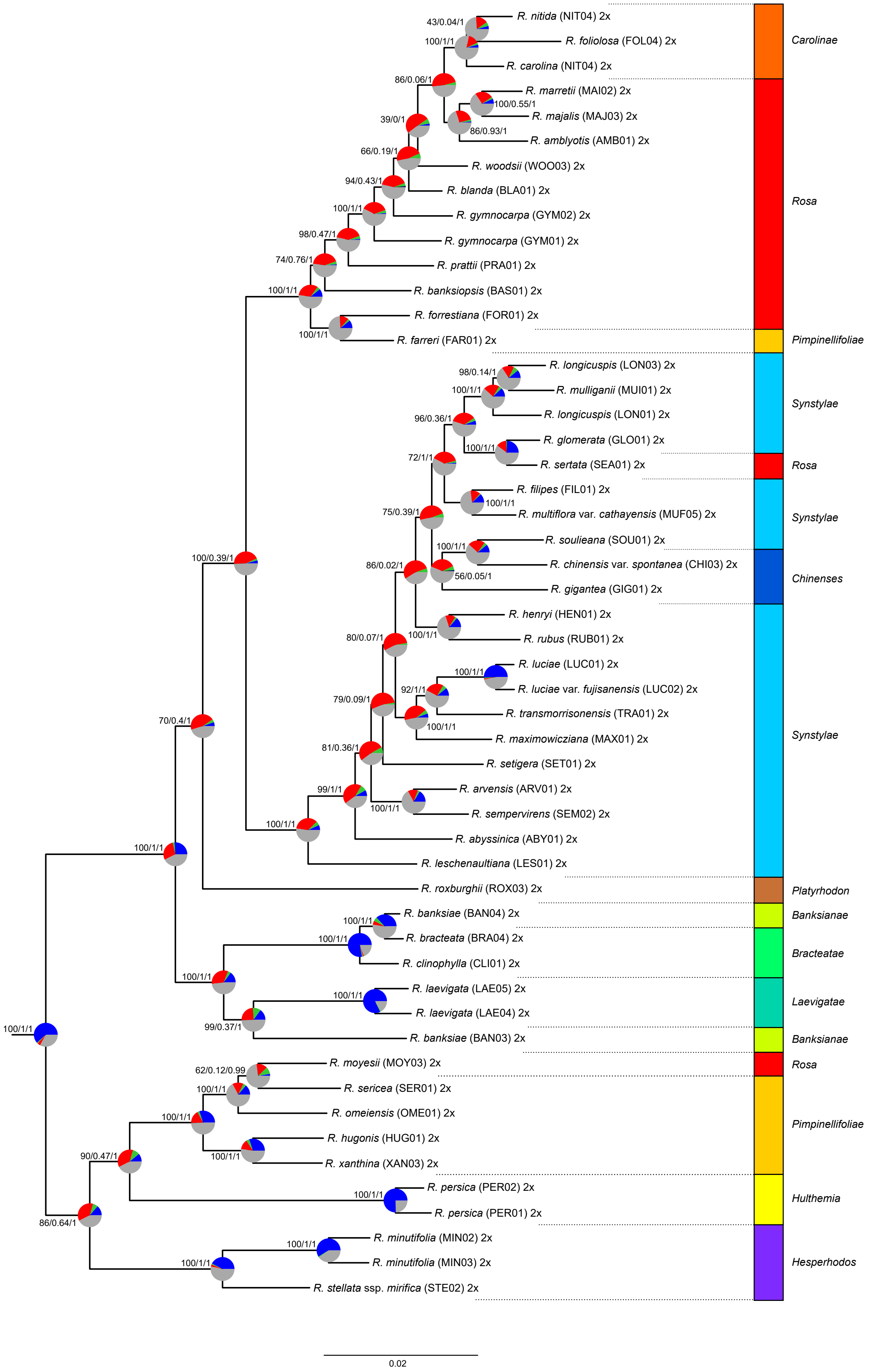


(Caption next page.)

Figure 34: (Previous page.) Phylogenetic relationships among *Rosa* species as reconstructed using plastid $\mathrm{SCO}_{\mathrm{Tag}}$s. The topology presented corresponds to the best ML tree with 1000 bootstrap replicates to assess the support of branches. Bootstrap supports are indicated above the subtending branch. * indicates a maximum bootstrap support (i.e., 100). Stars below branches indicate that the corresponding branch was retrieved with a posterior probability >0.90 in the Bayesian analysis (see Appendix B, Supplementary Figure B.6). Branches with less than 50% of bootstrap support were collapsed. Blue leaves highlight hybrid diploids and bold leaves designate polyploids. The ploidy level of each accession, as estimated at step 2 (Figure 32), is mentioned after its name and accession code. MC1 and MC2 refer to supported meta clades. C1 to C5 correspond to subclades or grades embedded in either MC1 or MC2. Outgroups are not shown.

there was generally little difference between likelihoods of each scenario (Table 4). We therefore considered the 16 accessions as diploid hybrids and removed them to draw the backbone phylogeny of diploid accessions.

For biallelic $\mathrm{SCO}_{\mathrm{Tag}}$s in non-hybrid diploids, one consensus allele sequence was produced, mixing the variants of each of the two copies. Four methods for phylogenetic reconstructions were then considered to develop a strong backbone phylogeny hypothesis that would represent bifurcating relationships between diploid non-hybrid progenitors. Each method differs in its approach to phylogenetic reconstruction but, altogether, they provide an overall view of the support of the backbone phylogeny. Two methods used a concatenation of all consensus alleles (ML and Bayesian inferences), whereas the two others (coalescent tree and conflict study) build their phylogenetic hypotheses based on individual consensus gene trees. The backbone phylogeny shows well-resolved branches in concatenation-based methods that often reached the maximal BS or posterior probability (Figure 35). Most recent nodes remain unchanged between the plastid phylogeny and the backbone phylogeny. We encountered the inclusion of *R.* sect. *Carolinae* in *R.* sect. *Rosa*, and *R.* sect. *Chinenses* in *R.* sect. *Synstylae* again. However, the organization of deep nodes changed between the plastid and the nuclear backbone phylogenies. *R.* sect. *Pimpinellifoliae* branched at the base of the genus together with the two subgenera *Hesperhodos* and *Hulthemia*. The remaining accessions form three clades: a *Rosa*-*Carolinae* clade, a *Synstylae*-*Chinenses* clade and a *Banksianae*-*Bracteatae*-*Laevigatae* clade. The first two clades present more topological conflict than the other clades, which is revealed by the higher number of alternative bipartitions and the lower LPP values. In detail, most of the accessions fall into clades that approximately correspond to their section, except for a few individuals. *R. farreri* (FAR01), *R. banksiae* (BAN04) and *R. moyesii* (MOY03) fill unexpected positions but their placements are similar to those observed in the plastid phylogeny, which suggests either misidentification of the sample or wrong assignment of the species to its section. *R. abyssinica* (ABY01), *R. sertata* (SEA01), *R. setigera* (SET01) and *R. glomerata* (GLO01) show contradictory placements between the plastid phylogeny and the nuclear backbone phylogeny. This may highlight hybrid specimens that were not detected in our step of diploid hybrid identification due to either insufficient allele coverage or variation.



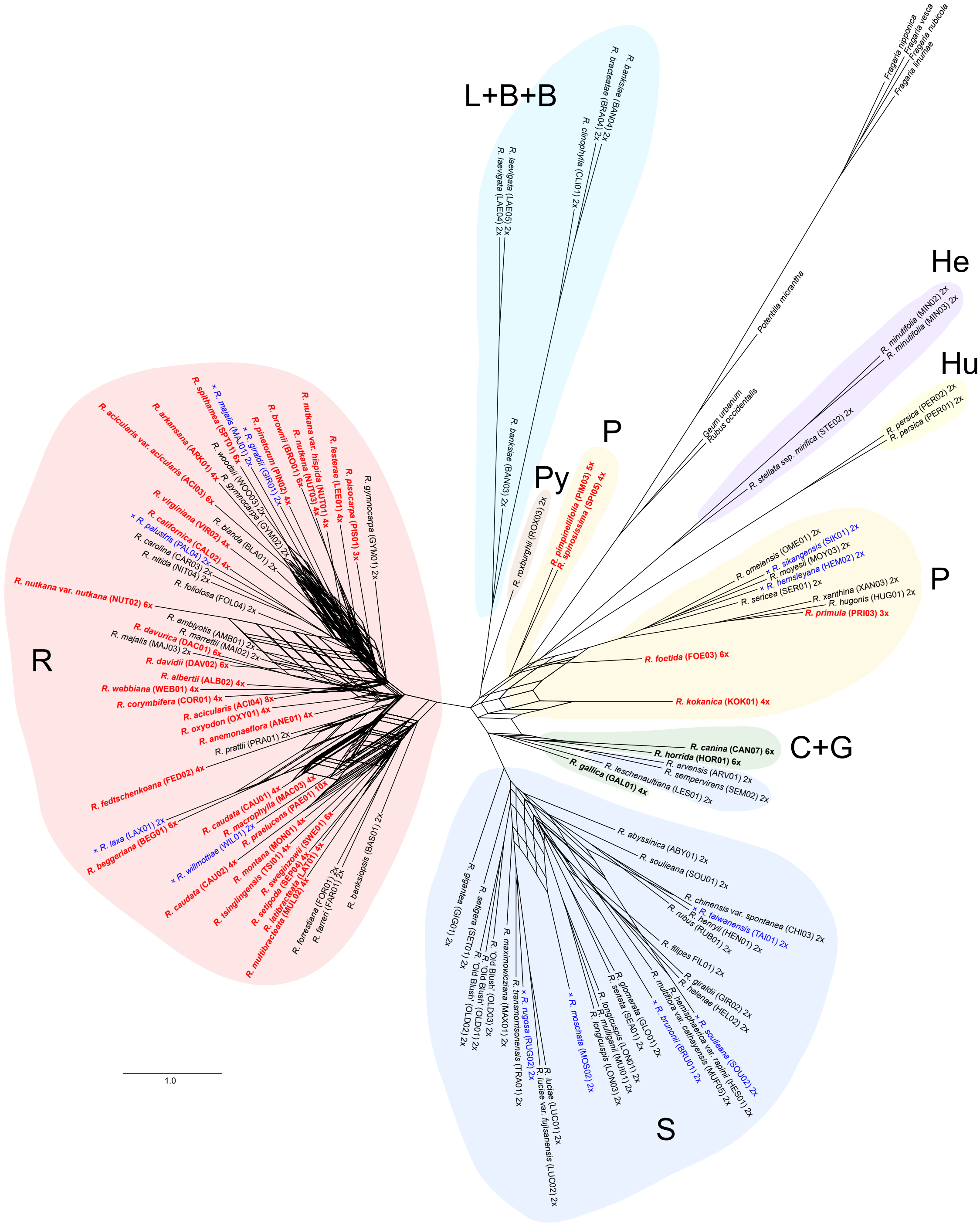
(Caption next page.)

Figure 35: (Previous page.) Backbone phylogeny of non-hybrid diploid putative progenitors of *Rosa*. The topology and branch lengths are from the best ML tree. The same topology was recovered in the Bayesian analysis. Leaf labels correspond to the species' name, its accession code and its ploidy level, as estimated at step 2 (Figure 32). The three slash-separated numbers indicate the bootstrap support (1000 replicates) from the ML analysis, the local posterior probability from Astral, and the posterior probability from the Bayesian analysis, respectively. The pie charts at each node present the fraction of SCO$_{\mathrm{Tag}}$s that supports that bipartition, the fraction that supports the main alternative bipartition (green), the fraction that supports other alternative bipartitions (red) and the fraction with either less than 50% of bootstrap support or that do not have this bipartition due to missing data (gray). Outgroups are not shown.

3.4.5 Placements of hybrid and polyploid taxa show multiple intra and intersectional reticulations

The backbone phylogeny provided a well-supported frame for the study of reticulations (Figure 35). We therefore wanted to study the placement of diploid hybrids and polyploids on this backbone phylogeny. In particular, some *Rosa* sections only consist of polyploid species (*R.* sect. *Caninae*, *R.* sect. *Gallicanae*) and were therefore not present in the backbone phylogeny. The split network, obtained after transformation of a MUL tree, itself produced by coalescence of allele trees, provides an overall view of the placement of diploid hybrids and polyploids. To ease the representation and interpretation of split networks, we split the genus according to the meta clades MC1 and MC2 identified on the plastid phylogeny. Most polyploids are located within *R.* sect. *Rosa*, *Carolinae*, *Caninae*, *Gallicanae*. The split network related to MC1 (Figure 36) shows that all polyploid accessions from *R.* sect. *Rosa* and *Carolinae* fall with their respective diploid counterparts into a distinct group. *R. pimpinellifolia* (PIM03) and *R. spinosissima* (SPI05), both *Pimpinellifoliae* species, have an intermediate position between the core *Pimpinellifoliae* group and the *Rosa*-*Carolinae* group. While *R. praelucens* (PAE01) is assumed to belong to the subgenus *Platyrhodon*, it is found here to be a sister to Asian accessions from *R.* sect. *Rosa*. The split network related to MC2 (Figure 37) identifies a group including *Caninae*-*Gallicanae* accessions that is situated in between the *Rosa*-*Carolinae* group and the *Synstylae*-*Chinenses* group. The *Synstylae* group is split, with the European accessions on one side and the remaining Asian accessions on the other side. European *Synstylae* branch at the base of the *Caninae* group.

Using a combination of observations of the split networks and previous hypotheses, we further investigated putative hybridizations. Due to computing resource limitations, we only focused on 12 hybridization scenarios that we studied in a maximum pseudo likelihood (MPL) framework that is also able to provide the contribution of each parental lineage to the creation of the hybrid specimen through inheritance probabilities (Figure 38). We confirm the intersectional hybrid origin of *R.* sect. *Caninae* where most of the species are pentaploid (5x). Around 60% of their genome may have been inherited from European *Synstylae* species, while 40% may come from Eurasian *Rosa* accessions. *R. marginata* may have resulted from a cross between *Caninae* lineages and lineages



(Caption next page.)

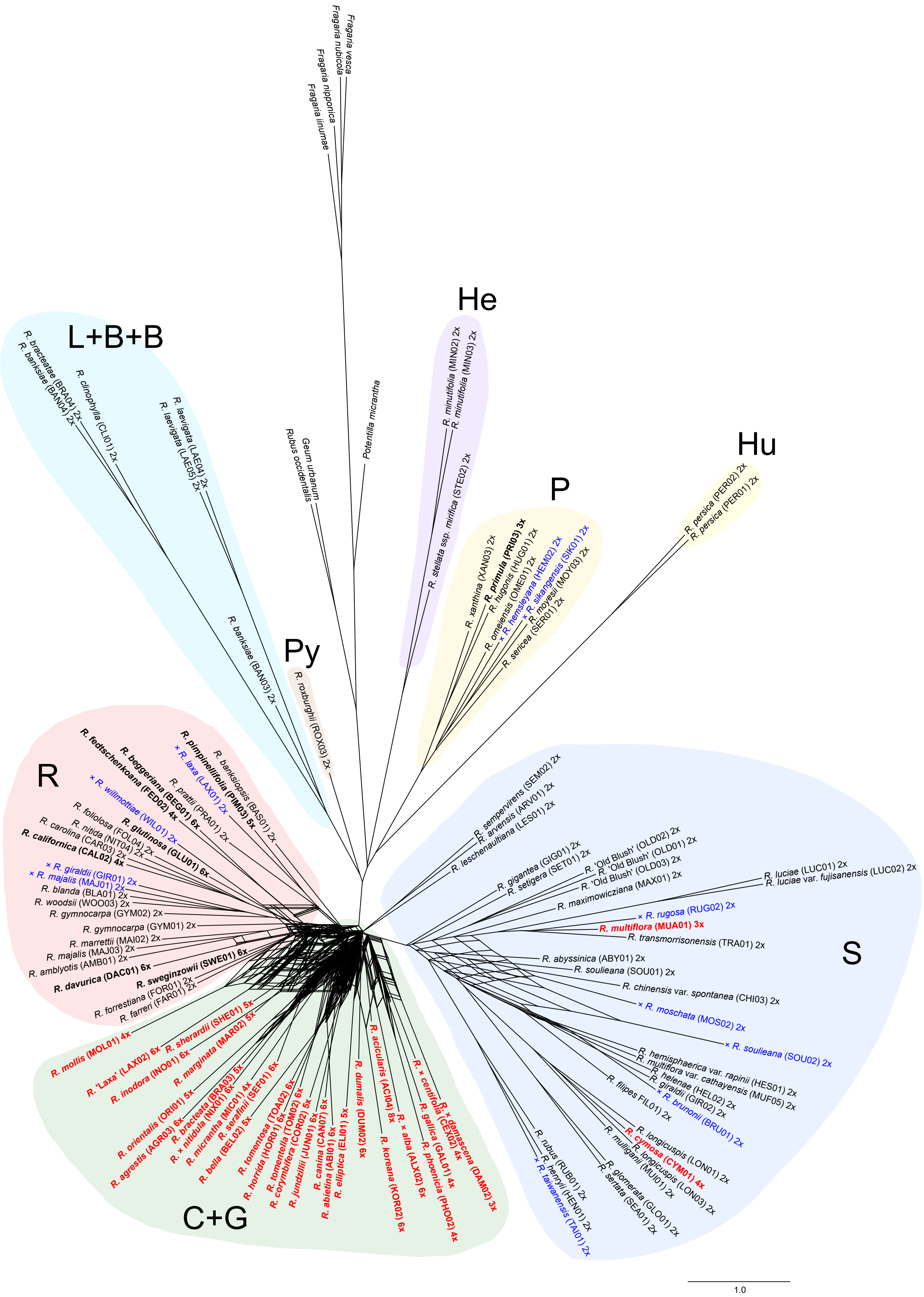
Figure 36: (Previous page.) Split network representation of the placement of hybrid diploids and polyploids of meta clade 1. Hybrid diploids from MC1 and MC2 are in blue. Polyploids from MC1 are in bold and red. Polyploids from MC2 are in bold and black and serve as placeholders for larger clades. Diploids from the backbone phylogeny are in black. Leaf labels correspond to the species' name, its accession code and its ploidy level, as estimated at step 2 (Figure 32). Groups are named according to the main sections/subgenenus dominating each group. C+G: *Caninae* and *Gallicanae*; He: *Hesperhodos*; Hu: *Hulthemia*; L+B+B: *Laevigatae*, *Banksianae*, *Bracteatae*; P: *Pimpinellifoliae*; Py: *Platyrhodon*; R: *Rosa*; S: *Synstylae* and *Chinenses*.

from section *Rosa*. In the same way, *R. pimpinellifolia*, usually considered to be a tetraploid, may be a 50-50 intersectional hybrid between Middle Eastern lineages of section *Rosa* (0.53 ~1/2) and a yellow-flowered lineage of *R.* sect. *Pimpinellifoliae* (0.47 ~1/2). A similar observation can be made for *R. foetida*, which may be derived from a hybridization between the main *Pimpinellifoliae* clade and lineages within the section *Rosa*, but inheritance probabilities are less clear (0.58 and 0.42). Regarding *Gallicanae* species, *R. gallica* (4x) may originate from an extinct lineage at the base of the *Synstylae*-*Caninae* clade (0.46 ~1/2), and the European *Synstylae* clade (0.54 ~1/2). *R.* × *damascena* (3x) seems to be an intersectional hybrid between *R. gallica* (0.69 ~2/3) and the *Synstylae R. brunonii* (0.31 ~1/3). *R. moschata*, found to be a diploid hybrid in our study, may be the result of a cross between two *Synstylae* lineages (East Asian *Synstylae R. luciae* (0.52 ~1/2) × West Asian *Synstylae R. brunonii* (0.48 ~1/2)). Regarding North American species from sections *Rosa* and *Carolinae*, tetraploid *R. californica* and *R. virginiana* seem to be a mix of lineages from the two sections but with inheritance probabilities that do not clearly match their ploidy level, once again showing their intertwined relationships. We confirm the placement of *R. praelucens*, found near Asian species from section *Rosa*, which may derive from *R. sweginzowii* (0.77), also found at high altitudes and a deep ancestor of Asian diploid lineages (0.23). We considered that *R. cymosa* (CYM01) was tetraploid but found no clear evidence for an allopolyploid origin in the hybrid network (92% inherited from a *Synstylae* lineage and 8% from a *Chinenses* lineage). In addition, this accession holds an odd position since it should be grouped with other Banksianae accessions. Both misidentification of the sample and over-estimation of the ploidy level may explain (1) the inconsistencies observed between the inheritance probabilities, and (2) the odd positioning of the accession.

3.5 Discussion

3.5.1 The development of a framework for the study of complex taxonomic entities

Despite the fact that polyploidization is actually known as a major diversification force in plants, most previous phylogenetic studies considered the relationships between taxa through the lens of only a few markers, usually from the plastid genome, which are easier to target than nuclear



(Caption next page.)

Figure 37: (Previous page.) Split network representation of the placement of hybrid diploids and polyploids of meta clade 2. Hybrid diploids from MC1 and MC2 are in blue. Polyploids from MC2 are in bold and red. Polyploids from MC1 are in bold and black and serve as placeholders for larger clades. Diploids from the backbone phylogeny are in black. Leaf labels correspond to the species' name, its accession code and its ploidy level, as estimated at step 2 (Figure 32). Groups are named according to the main sections/subgenenus dominating each group. C+G: *Caninae* and *Gallicanae*; He: *Hesperhodos*; Hu: *Hulthemia*; L+B+B: *Laevigatae*, *Banksianae*, *Bracteatae*; P: *Pimpinellifoliae*; Py: *Platyrhodon*; R: *Rosa*; S: *Synstylae* and *Chinenses*.

DNA. While these sequences significantly contributed to a better knowledge of plant phylogeny and made it possible to resolve many clades, they bear significant flows such as unilateral inheritance, haploid state and a slower rate of evolution, making them insufficient to study complex evolutionary processes such as hybridization and polyploidization. While reticulate evolutions are now increasingly taken into account in phylogenetic studies, the lack of general models and the computational burden associated with the treatment of large arrays of melted sequences restrict their study to either groups with few taxa (Kamneva et al., 2017) or to experimental designs involving few markers (Marcussen et al., 2012, 2015; Cai et al., 2012). By taking the genus *Rosa* as an example of complex taxonomic entity, we developed a general framework for taking hybridization and polyploidization into account in large taxonomic complexes. The stepwise framework that we developed has the advantage of starting without too much bias on the sampling because we estimated the ploidy level and the hybrid status of each sample during the analysis rather than relying solely on the literature or herbarium vouchers. We were therefore able to identify odd samples whose placements could be explained by either hybridization or misidentification. We adopted a diploids-first strategy, which consists of first identifying non-hybrid diploid species to draw a backbone phylogeny that would represent putative lineages for more complex species (diploid hybrids and polyploids) that were further grafted onto the diploid skeleton. Diploids-first approaches have already proved to be useful in other challenging taxonomic groups such as ferns (Beck et al., 2010), grass (Díaz-Pérez et al., 2018) and other Rosaceae lineages (Lo et al., 2010; Burgess et al., 2015). The diploids-first strategy obviously requires knowledge of the ploidy level of each accession as well as its hybrid status. We showed that allele frequencies at heterozygous SNPs make it possible to easily separate diploids from polyploids as long as there is sufficient sequencing coverage, a coverage that could be obtained with Illumina amplicon sequencing. The amplicon sequencing technique successfully targeted allele sequences at each locus and therefore collected allele variations from both parental lineages. We used these allele variations within each diploid accession to detect hybridization patterns, and we first identified hybrid diploids based on allele splitting in allele tree topologies. Hybrid candidates were further tested in a ML framework and, despite few variations between the likelihood of hybrid scenarios, most of the hybrid specimens were confirmed (Table 4). However, cross comparisons between the plastid phylogeny and the nuclear backbone phylogeny revealed additional patterns of putative hybridizations that were not covered when nuclear allele tree topologies were scanned alone. This highlights the importance

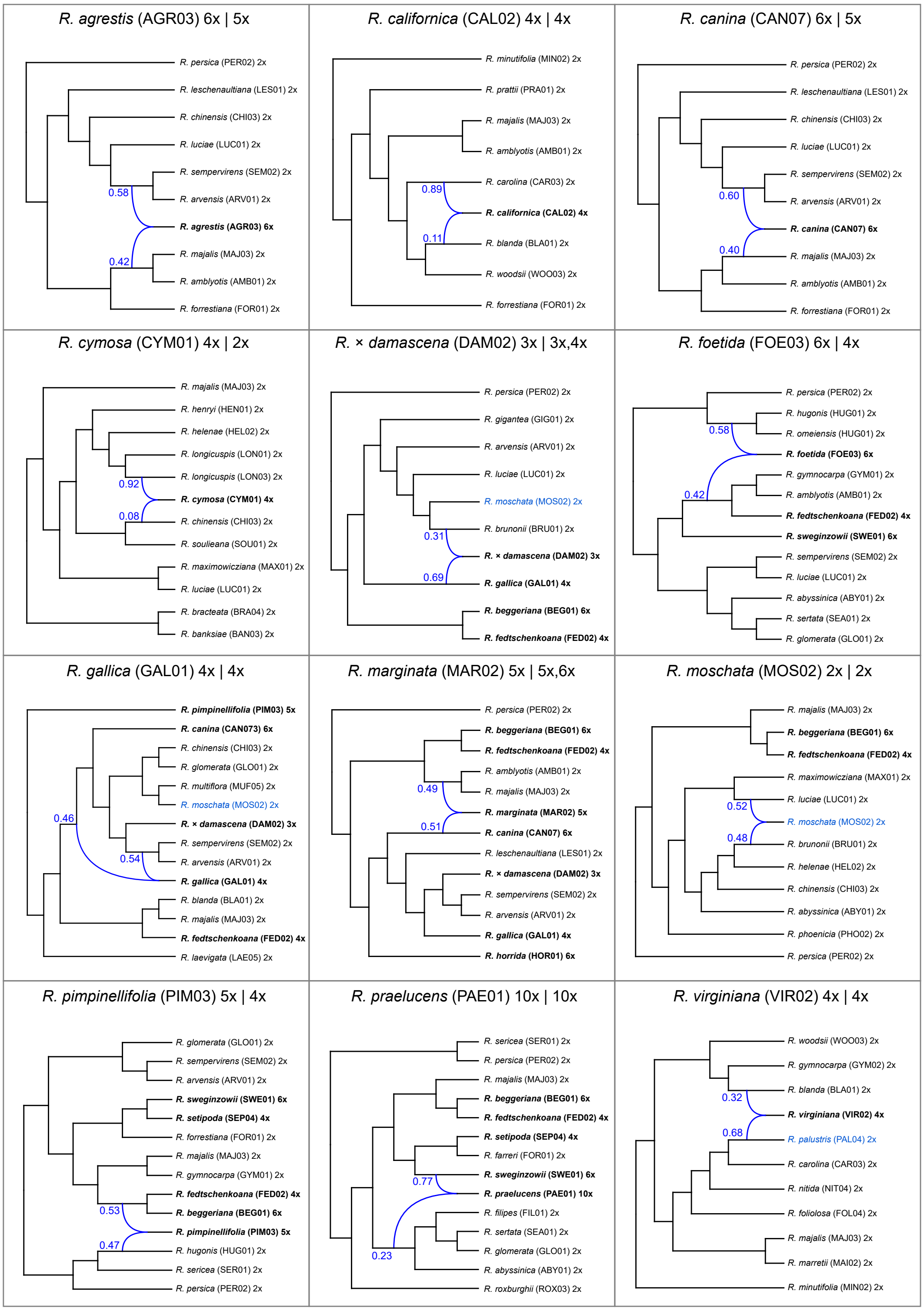


Figure 38: Hybrid origins of some *Rosa* specimens. Maximum pseudo likelihood networks show the hybrid origin of *Rosa* samples and the genome contribution of each parental lineage. Leaf labels correspond to the species' name, its accession code and its ploidy level, as estimated at step 2 (Figure 32). The ploidy levels indicated after the vertical bar correspond to the levels commonly reported for the species in the literature. The names of polyploids are in bold. The names of diploid hybrids are in blue.

.

of studying both plastid and nuclear data in phylogenomic analyses. We devoted much effort to building a robust plastid phylogeny by selecting the most informative regions of the chloroplast genome and taking INDEL variations into account. The resulting plastid phylogeny helped us to gain knowledge about our sampling and to pinpoint sections and accessions that would be interesting to further study reticulations. After obtaining a global view of the placement of hybrid and polyploid accessions through split networks, we selected some accessions for an in-depth study of their reticulation pattern and we were able to confirm most of our hybrid hypotheses and to provide more detail about the fractions inherited from the two parental lineages.

3.5.2 Toward a revision of the classification of the genus *Rosa*

Our stepwise study allowed us: (1) to obtain the most robust plastid and nuclear phylogenies of the genus *Rosa* to date; and (2) to shed light on several patterns of hybridization within the genus *Rosa*. Several hypotheses were previously made regarding reticulations in the genus *Rosa* and most of them were confirmed and deepened in our analysis. In addition, we shed new light on intersectional hybridizations that were never considered before.

In the currently used system of Rehder (1940), the genus *Rosa* is divided into four subgenera with *R.* subg. *Rosa* being further divided into ten sections. We confirmed the results of previous analyses (Bruneau et al., 2007; Fougère-Danezan et al., 2015) where *R.* subg. *Rosa* is not monophyletic and other subgenera do not branch at the base of the phylogeny. These observations support the treatment of *R.* subg. *Platyrhodon*, *Hulthemia* and *Hesperhodos* at the sectional level. The sections *Carolinae* and *Chinenses* are embedded in sections *Rosa* and *Synstylae*, respectively, in both plastid and nuclear analyses. We therefore suggest that *R.* sect. *Carolinae* is merged with *R.* sect. *Rosa* and that *R.* sect. *Chinenses* is merged with *R.* sect. *Synstylae*, in line with previous observations and suggestions (Fougère-Danezan et al., 2015). We also report an allopolyploid origin of *R. praelucens*, as suggested by Fougère-Danezan et al. (2015). However, the species is mostly derived from a cross between *R.* sect. *Rosa* lineages and we did not demonstrate any relationship between *R. praelucens* and its putative diploid counterpart *R. roxburghii*. We therefore recommend considering *R. praelucens* as a full member of the *Rosa* section and letting *R. roxburghii* be the representative species of the *R.* subg. *Platyrhodon*.

We also support several intersectional hybridizations that reveal the need to consider the genus *Rosa* as a hybrid system. We demonstrate the hybrid origin of *R. spinosissima* (syn. *R. pimpinellifolia*) that derives from a cross between *R.* sect. *Rosa* and *R.* sect. *Pimpinellifoliae*, as suggested decades ago based on karyotyping (Hurst, 1928), and more recently in Fougère-Danezan et al. (2015). This may explain the hardiness of this species, which covers a huge area of distribution and thrives in very different ecosystems, from the British coasts to the Altai Mountains. We also demonstrate the origin of some individuals of the *Caninae* section that derived from European diploid lineages of *R.* sect. *Synstylae* for three-fifths and *R.* sect. *Rosa* for two-fifths. *R. marginata*, previously assumed to be a hybrid between *R. gallica* and a species from *R.* sect. *Caninae* (Wissemann, 1999), is actually best resolved as an intersectional hybrid between *R.* sect. *Rosa* and *R.*

sect. Caninae. *R. gallica* derives from an extinct lineage at the base of the *Synstylae*-*Caninae* clade and European *Synstylae* species. Finally, species that are controversial regarding their wild origin, such as *R.* × *damascena* and *R. moschata*, were found to be hybrids between lineages that may have been selected by men for scent-related traits. We confirm the contribution of *R. moschata*/*R. brunonii* and *R. gallica* to the formation of *R.* × *damascena*, with *R. moschata*/*R. brunonii* being the maternal lineage, as suggested by our plastid analysis, in line with observations made by Iwata et al. (2000). However, unlike Iwata et al. (2000), our *R.* × *damascena* accession was found to be triploid and we did not find any phylogenetic relationships between *R.* × *damascena* and the section *Rosa*.

3.6 Conclusion

Recent advances in sequencing techniques currently make it possible to target not only plastid sequences but allele variations within each individual as well. We took advantage of this wealth of information to study patterns of reticulation in the genus *Rosa*. We developed a stepwise diploids-first strategy that successfully recovered robust phylogenies and provided unprecedented insights into the many patterns of hybridization occurring in this genus. We resolved most of the deep branches of the *Rosa* phylogeny that mainly concern phylogenetic relationships between subgenera and sections. These robust phylogenies as well as the supported patterns of hybridization provide the ground to further revise the nomenclature of the genus *Rosa* in order to integrate the hybrid dimension of certain sections. Our robust plastid and nuclear phylogenies also provide a well-supported framework for studying the evolution of ornamental traits in *Rosa*. Relationships between species belonging to large sections still remain unclear and we suggest using concepts from population genetics on a broader sampling to uncover intrasectional species relationships.

Several groups of plants exist in which phylogenies remain unclear due to complex evolutionary processes. Such processes may correspond to intricate cytological events with whole genome duplications that lead to variations in chromosome numbers (*Lobelia*, Bambusae). Moreover, the evolutionary history of a group is sometimes further complicated by the hand of man in the case of agricultural or ornamental interests (Pyrinae, *Cotoneaster*, *Rubus*, *Hieracium*). By addressing phylogenetic relationships between *Rosa* species, we took a step forward in the analysis of reticulate evolutions on large and complex taxonomic entities and we hope to have provided a framework that could be reproduced in other challenging groups.

3.7 Additional file and availability of data and material

Silica dried material present in Additional File 1 are available upon request and are subject to availability. Primer pairs information associated with the 96 SCO_{Tag}s are available in Additional File 2. All the filtered SCO_{Tag} alleles are available in Additional File 3. Some custom scripts developed for this study are available at `https://github.com/kdebray/RosaPhylogenomics`. The entire set of Illumina paired-end read sequences have been deposited at DDBJ/ENA/GenBank under the accession PRJNA591118.

Additional File 1. Vouchers of all specimens with tissue material and/or DNA extract preserved at IRHS (Excel file).

Additional File 2. Genomic localization and primer pair information associated with the 96 $\mathrm{SCO}_{\mathrm{Tag}}$ sequences used in the analysis (Excel file). Chromosome names and $\mathrm{SCO}_{\mathrm{Tag}}$ positions are based on the genome sequence of Hibrand Saint-Oyant et al. (2018).

Additional File 3. Filtered alleles from all accessions at all $\mathrm{SCO}_{\mathrm{Tag}}$ loci after the read processing and assembly steps (Fasta file).

The two aforementioned Additional Files and sample material are available upon request to the GDO team.

3.8 Author's contributions

KD, FF, VM conceived and designed the study. KD and VM sampled some of the wild *Rosa* accessions present in the collection of tissue material. KD extracted DNA and designed the $\mathrm{SCO}_{\mathrm{Tag}}$ sequences. MCLP and AB conducted and managed the amplicon production and sequencing. KD designed the bioinformatics pipeline for amplicon sequencing analysis and performed the phylogenetic studies. KD drafted the manuscript and FF and VM proofread the manuscript. JMM assembled the plastid genome sequences of *R. gallica*, *R. laevigata*, *R.* 'Old Blush', *R. persica*, *R. xanthina* and KD assembled the plastid genome sequences of *R. arvensis*, *R. canina*, *R.* × *damascena*, *R. dumalis*, *R. elliptica*, *R. gigantea*, *R.* × *odorata* and *R. wichurana*. TT and GM helped with sampling *Rosa* accessions in rose gardens.

3.9 Acknowledgments

The authors thank A. Bruneau (University of Montréal) for providing samples of *Rosa* used in her past studies. We also thank the Arboretum du Vallon de l'Aubonne, the Arnold Arboretum of Harvard University, the Botanical Garden of the University of Bern, the Berlin-Dahlem Botanical Garden and Botanical Museum, the Botanical Garden of the University of Bonn, the Cambridge University Botanic Garden, the Conservatoire Botanique National de Franche-Comté, the Botanical Garden of the University of Strasbourg, the Botanical Garden of the City of Nantes, the Kalmthout Arboretum, Les Chemins de la Rose rose garden, the Botanical Garden of the Finnish

Museum of Natural History, the Maribor University Botanic Garden, the Ferdowsi University of Mashhad, the Memorial University of Newfoundland, the Meise Botanic Garden, the Morton Arboretum, the Botanical Garden of the University of Padua, the Quarryhill Botanical Garden, the Royal Botanic Garden of Edinburgh, the Roseraie du Val-de-Marne, Roses Loubert, the Rancho Santa Ana Botanic Garden, the University of California at Davis, the Unité Expérimentale de la Villa Thuret and Wageningen University & Research for preparing and/or providing *Rosa* samples. The authors acknowledge Isabelle Le Clainche and Aurélie Chauveau, Aurélie Canaguier and Elodie Marquand from the EPGV group for wet-lab amplicon production and sequencing, data transfer and public data submission to NCBI. The authors are grateful to the Genotoul Bioinformatics Platform of Toulouse Midi-Pyrenées (Bioinfo GenoToul) for providing computing and storage resources and bioinformatics support. The authors also thank Gail Wagman for reviewing the English of the manuscript.

3.10 Supplementary results Part A: Phylogenetic analysis of the genus *Rosa* using chloroplast whole genome sequences

3.10.1 Introduction

Chloroplasts in green plants' cells are considered endosymbiotic *Cyanobacteria* that were engulfed in a eukaryotic cell (Whatley et al., 1979; Cavalier-Smith, 1982; Douglas, 1998) around 1 to 2 BYa (McFadden and van Dooren, 2004; Sánchez-Baracaldo et al., 2017). Chloroplasts have their own DNA sequence that is mostly circular and haploid, and distinct from the nuclear genome, although there exist some genomic transfers between the two cellular components (Rousseau-Gueutin et al., 2018). As for the mitochondrial genome, the chloroplast genome is generally inherited from one parents in angiosperms (Reboud and Zeyl, 1994; Birky, 1995). In *Rosa*, Corriveau and Coleman (1988) demonstrated that plastid DNA is maternally inherited in *Rosa rugosa*, and this result may be extrapolated to the whole genus. Using F1 progeny, maternal inheritance was also demonstrated in cultivated roses as 'Old Blush' (unpublished results from the IRHS-GDO team). The rate of substitution in chloroplast DNA sequences are generally lower than those observed in the nuclear genome (Wolfe et al., 1987). Chloroplast sequences are therefore more conserved in land plants (Wicke et al., 2011; Dong et al., 2013). The circular chloroplast genome is typically 120,000-170,000 bp long in land plants (Jiao and Guo, 2014) and displays a four-part structure with two inverted repeats (IRa and IRb) that separate a long single-copy section (LSC) from a short single-copy region (SSC) (Shaw et al., 2007). In *Rosa*, the chloroplast genome is on average 156,500 bp long. Plastid sequences were extensively used to address the phylogenetic relationships between species of the genus *Rosa* (Wissemann and Ritz, 2005; Bruneau et al., 2007; Qiu et al., 2012, 2013; Fougère-Danezan et al., 2015). Despite the fact that plastid sequences cannot trace reticulate relationships per se due to their unilateral inheritance, they still represent a valuable resource for phylogenetics since they are easily targeted and avoid to consider allelic data. However, due to their high level of conservation, plastid sequences have often difficulty to resolve recent speciation events, generally occurring at the genus level, compared to single-copy nuclear genes such as *Adh* or *agt1* (Sang, 2002; Naumann et al., 2011). Resulting phylogenies that rely on few plastid sequences often lack support at recent times or at times of the evolutionary history that correspond to rapid radiations (Bruneau et al., 2007; Fougère-Danezan et al., 2015). To circumvent this aspect, one solution would be to compare chloroplast whole genome sequences to gather sufficient variations in order to unravel species relationships.

Many studies now resort to whole genome sequencing to address various scientific questions in biology, from medicine to environment. Therefore, public sequence databases are plenty of untapped datasets that contain informative variations for phylogenetics (Wuyts and Segata, 2019). Due to the large presence of chloroplasts in green plants' cells (estimated between 80-120/mesophyll cell (Crumpton-Taylor et al., 2012)), genomic sequencing based on photosynthetic material may include a substantial fraction (5-10%) of reads that originate from the chloroplast genome, even if nuclear purification is regularly done (Ahmed, 2015). Plastid genomes are therefore often well covered even in sequencing project performed at low depths (5-30X). At the same time, many software were recently developed to easily assemble plastid genomes based on genome skimming data (Dierckxsens et al., 2017; Al-Nakeeb et al., 2017; Jin et al., 2018; Meng et al., 2019).

Here, we searched public databases for the presence of chloroplastic DNA in sequencing read sets of *Rosa* species. Genomic reads were used to assemble chloroplast whole genome sequences that further served for a phylogenetic study.

3.10.2 Materials and methods

Nine chloroplast genome sequences were previously assembled by Jordan Marie-Magdelaine at IRHS (unpublished results of the IRHS-GDO team). Sixteen already assembled chloroplast whole genome sequences were retrieved from Genbank. Twenty additional chloroplast whole genome sequences were assembled from raw reads in these complementary results. The total number of *Rosa* accessions used in this study is 45 (Table 5).

For the assembly of chloroplast genome sequence not yet assembled (Table 5, b-labeled species names), genomic DNA sequencing reads were retrieved from Genbank using the identifiers provided in Table 5. We then ran NOVOPlasty2.7.1 (Dierckxsens et al., 2017) to target the assembly of the chloroplast genome sequence of each accession using *R.* chinensis 'Old Blush' (SRX3850987) as a reference (Hibrand Saint-Oyant et al., 2018). When the sequencing was performed at low depths and that the specimen is phylogenetically distant to *R.* 'Old Blush', we used the chloroplast genome sequence of a closer species to maximize the chance to recover a complete chloroplast genome sequence.

Chloroplast whole genome sequences were then aligned with Mafft v7.313 (Katoh and Standley, 2013). The orientations of LSC and SSC from already assembled public chloroplast genome

Table 5: The 45 *Rosa* accessions used for chloroplast whole genome phylogeny. a: Assembled by Jordan Marie-Magdelaine, b: Assembled by Kevin Debray, c: Already assembled and public available chloroplast whole genome sequences.

Species	Genbank ID	Origin	Publication
[c]*Rosa acicularis*	BOP011170	Genhe province, Inner Mongolia, China (121°36'1.00" E, 50°50'44.00" N)	Chen et al. (2019)
[b]*R. arvensis*	SRX3286288	Jardin expérimental de Colmar, France	Raymond et al. (2018)
[b]*R. arvensis*	SRX5082753	Botanical Garden of Würzburg, Germany	Unpublished
[c]*R. banksiae* var. *normalis*	MK361034	Midu county, Yunnan, China	Wang et al. (2019)
[c]*R. berberifolia*	MK423879	Manas County, Xinjiang Uygur Autonomous Region, China (44°06'13"N, 86°21'07"E, 864 m)	Zhang et al. (2019)
[b]*R. canina*	ERX1733250	GLM12396, Herbarium Senckenbergianum Görlitz (GLM), Germany	Unpublished
[b]*R. canina*	SRX5082754	Brno, Czech Republic (49.2310 N 16.5957 E)	Unpublished
[a]*R. chinensis* var. *spontanea*	SRX4006790	Roseraie du Val-de-Marne, L'Hay-les-Roses, France	Hibrand Saint-Oyant et al. (2018)
[a]*R. chinensis* 'Old Blush'	SRX3850987	Institut de Recherche en Horticulture et Semences, Angers, France	Hibrand Saint-Oyant et al. (2018)
[c]*R. chinensis* 'Old Blush'	MH332770	Flower Research Institute of Yunnan, Academy of Agricultural Sciences, Yunnan, China	Li et al. (2019)
[c]*R. chinensis* 'Old Blush'	PDCK01000046	École Normale Supérieure, Lyon, France	Raymond et al. (2018)
[b]*R. chinensis* var. *mutabilis*	SRX3286282	École Normale Supérieure, Lyon, France	Raymond et al. (2018)
[b]*R. chinensis* var. *sanguinea*	SRX3286285	École Normale Supérieure, Lyon, France	Raymond et al. (2018)
[c]*R. chinensis* var. *spontanea*	MG523859	Yichang, Hubei, China (111°10' E, 30°47' N, 400 m)	Jian et al. (2018)
[b]*R. chinensis* var. *spontanea*	SRX3286289	La Bonne Maison, La Mulatière, Lyon, France	Raymond et al. (2018)
[b]*R. corymbifera*	SRX5082752	Weienberg, Germany (51.1732 N 14.6271 E)	Unpublished
[b]*R.* × *damascena*	SRX3286290-SRX3286291	École Normale Supérieure, Lyon, France	Raymond et al. (2018)
[b]*R. dumalis*	ERX1733252	GLM49831, Herbarium Senckenbergianum Görlitz (GLM), Germany	Unpublished
[b]*R. elliptica* subsp. *inodora*	ERX1733251	GLM49596, Herbarium Senckenbergianum Görlitz (GLM), Germany	Unpublished
[a]*R. gallica*	SRX4006794	Roses Loubert rose garden, Les Rosiers-sur-Loire, France	Hibrand Saint-Oyant et al. (2018)
[b]*R. gigantea*	SRX3286283-SRX3286284	Lyon Botanical Garden, France	Raymond et al. (2018)
[a]*R. laevigata*	SRX4006792	Roseraie du Val-de-Marne, L'Hay-les-Roses, France	Hibrand Saint-Oyant et al. (2018)
[c]*R. lucieae*	MG727864	Seoguipo, Jeju, Korea	Jeon and Kim (2019)
[b]*R. majalis*	SRX3286287	École Normale Supérieure, Lyon, France	Raymond et al. (2018)
[b]*R. majalis*	SRX5231945	GLM172056, Herbarium Senckenbergianum Görlitz (GLM), Germany	Unpublished
[c]*R. maximowicziana*	MG727865	Hwaseong, Gyeonggi, Korea	Jeon and Kim (2019)
[a]*R. minutifolia* var. *alba*	SRX4006787	Roseraie du Val-de-Marne, L'Hay-les-Roses, France	Hibrand Saint-Oyant et al. (2018)
[a]*R. moschata*	SRX4006793	Roses Loubert rose garden, Les Rosiers-sur-Loire, France	Hibrand Saint-Oyant et al. (2018)
[b]*R. moschata*	SRX3286292	La Bonne Maison, La Mulatière, Lyon, France	Raymond et al. (2018)
[c]*R. multiflora*	MG727863	Namyangju, Gyeonggi, Korea	Jeon and Kim (2019)
[c]*R. multiflora*	MG893867	NA	Unpublished
[c]*R. odorata* var. *gigantea*	KF753637	Kunming Botanical Garden, Kunming Institute of Botany, China	Yang et al. (2014)
[b]*R.* × *odorata* 'Hume's Blush'	SRX3286293	Lyon Botanical Garden, France	Raymond et al. (2018)
[b]*R. pendulina*	SRX3286278	Lyon Botanical Garden, France	Raymond et al. (2018)
[a]*R. persica*	SRX4006789	Roses Loubert nurseries, Les-Rosiers-sur-Loire, France	Hibrand Saint-Oyant et al. (2018)
[c]*R. praelucens*	MG450565	Tanganpei Village, Shangri-La County, China (99°49.635'E, 27°32.278'N, 3248m)	Jian et al. (2018)
[c]*R. roxburghii*	PRJNA356521	Meitan, Guizhou, China (107°28'15.01" E, 27°46'51.41" N)	Wang et al. (2018)
[c]*R. roxburghii*	KX768420	NA	Unpublished
[a]*R. rugosa*	SRX4006791	Roseraie du Val-de-Marne, L'Hay-les-Roses, France	Hibrand Saint-Oyant et al. (2018)
[c]*R. rugosa*	MK986659	NA	Unpublished
[b]*R. rugosa*	SRX3286286	Lyon Botanical Garden, France	Raymond et al. (2018)
[c]*R. rugosa* var. *angusta*	MK947051	Hagampo coast, Wonbuk-myeon, Taean-gun, Chungcheongnam-do, Korea	Kim et al. (2019)
[b]*R. spinosissima*	SRX5231947	GLM172057, Herbarium Senckenbergianum Görlitz (GLM), Germany	Unpublished
[b]*R. wichurana*	SRX3286280-SRX3286281	École Normale Supérieure, Lyon, France	Raymond et al. (2018)
[a]*R. xanthina* var. *xanthina* f. *spontanea*	SRX4006788	Roses Loubert rose garden, Les Rosiers-sur-Loire, France	Hibrand Saint-Oyant et al. (2018)

Table 6: Variability content in the four plastid genome regions. LSC: Long single-copy region; IR: Inverted repeat; SSC: Small single-copy region; PIS: Parsimony informative sites (gap included); rPIS: Parsimony informative sites (gap excluded); VAR: Variable sites (gap included); rVAR: Variable sites (gap excluded).

Region	PIS	rPIS	Var	rVar
LSC	7.65%	1.79%	13.54%	2.43%
IRa	0.89%	0.28%	1.32%	0.45%
IRb	0.80%	0.26%	5.18%	0.41%
SSC	5.46%	2.11%	17.79%	3.05%

sequences were sometimes reverse complemented in Geneious v9.1.7 (https://www.geneious.com) to match the orientation of the reference chloroplast genome sequence of *R.* 'Old Blush'.

The phylogenetic analysis was performed with RAxML v8.2.11 (Stamatakis, 2014) using (1) 1000 bootstrap replicates to infer a bootstrap support (BS) for each branch (2) a partitioning scheme in four subsets that corresponded to the four sub-regions (IRa, IRb, SSC and LSC). A GTR+G substitution model was applied to each subset of the partition . The chloroplastic genome of *Fragaria vesca* subsp. *vesca* (JF345175) was used as the outgroup species to root the phylogenetic tree.

3.10.3 Results

The chloroplast genome sizes ranged between 156,143 bp for *R. persica* (SRX4006789) to 158,565 bp for *R. majalis* (SRX5231945) (Additional file S1). The variability content differed from one region to another (Table 6). Inverted repeats contains 6 to 7 times less parsimony informative sites (gap excluded) than the LSC and the SSC, respectively. The most variable region in terms of parsimony informative sites corresponds to the SSC.

The phylogenetic tree based on chloroplast whole genome sequence is presented in Figure 39. Most of the branches are well supported and receive a maximal bootstrap support (100). However, we observed that two deep branches have support values of 74 and 71. These branches correspond to the separation of *R.* subg. *Hulthemia*, and subg. *Hesperhodos* versus the rest of the genus and to the separation of meta clade 1 (MC1) (Mostly accessions from *R.* sect. *Rosa*) and meta clade 2 (MC2) (Mostly accessions from *Synstylae* and *Caninae*). Most of the time, species are found close to their consectional relatives, except that *R. minutifolia* var. *alba* (SRX4006787) and *R. moschata* (SRX4006793) occupy odd places within MC1. *R.* sect. *Pimpinellifoliae* seems to diverge first in the evolutionary history of the genus, except that *R. spinosissima* is embedded in MC1. Known consectional polyploid species are generally located close to each other, such as the *Caninae* accessions. The subgenus *Platyrhodon* is split between *R. praelucens* in MC1 and *R. roxburghii* in MC2. *R. roxburghii* is at the basis of MC2, along with mono- bi-specific species sections *Banksianae* and *Laevigatae*, although their corresponding branches are weakly supported (BS<70%). *R.* × *damascena* is found closer to *R. moschata* than to *R. gallica* although both *R.* × *damascena* and *R. gallica* belong to section *Gallicanae*. All accessions of *R. chinensis* 'Old Blush' are closer to *R.* chinensis var. *spontanea* than to other varieties of *R. chinensis*. In addition, our results show that *R.* × *odorata* 'Hume's Blush' is somewhat closer to *R. chinensis* 'Old Blush' than to *R. odorata* var. *gigantea*. *R. arvensis*, a European *Synstylae* species, is sister to accessions from the European section *Caninae*.

3.10.4 Discussion

The phylogenetic relationships of *Rosa* species as inferred with chloroplast whole genome sequences provided similar results to those obtained in the previous chapter. Especially, we confirmed that section *Pimpinellifoliae* corresponds to an early lineage in the genus, and that subgenera and sections are largely embedded within each other. The two usual groups (MC1 and MC2) were found consistent with previous observations in Chapter 3. The lower support received by the subtending branch of MC1 and MC2 may reflect the rapid radiations that may have occurred after the steep temperature decrease concomitant with the *Grande Coupure* (around 34 MYa, see Fougère-Danezan et al. (2015)). Both *R.* × *damascena* and *R. spinosissima* are found far from their respective consectional relatives. These observations were already done in the previous chapter with other accessions from these two species. The odd positioning of these two species was attributed to their intersectional hybrid origin. Two accessions (*R. minutifolia* var. *alba* (SRX4006787) and *R. moschata* (SRX4006793)) seem to be misplaced with no particular biological reason, suggesting a misidentification of the samples. *R. minutifolia* var. *alba* (*R.* subg. *Hesperhodos*) should branch earlier in the phylogeny, together with section *Pimpinellifoliae* and subgenus *Hulthemia* (Fougère-Danezan et al., 2015) and Chapter 3. This discrepancy was also observed with nuclear data in Chapter 2 using the same accession. We therefore question the wild origin of this *R. minutifolia* var. *alba* accession. More field work should be carried out to verify the existence of *R. minutifolia* var. *alba* in the wild, as well as to make sure that this variety is assignable to *R. minutifolia* Engelm.. We also advocate that a resequencing of a true-to-type material of *R. minutifolia* Engelm. should be considered to include the genomic diversity of subgenus *Hesperhodos* in *Rosa* studies. As for *R. moschata* (SRX4006793), we had some doubts concerning its placement in Chapter 2. Here, we confirm its odd position, somewhat distant to its sister taxa (*Synstylae* clade), and raise the hypothesis that this garden accession has been misidentified.

3.10.5 Conclusion

The tremendous number of chloroplast whole genome sequences that were released over the past months provided 45 sequences to perform a phylogenetic study of *Rosa* at the chloroplast whole genome scale. We demonstrated that chloroplast whole genome sequences are easily assembled from raw genomic reads available on Genbank. The resulting phylogeny was highly supported and largely agrees with the plastid phylogeny obtained in Chapter 3. The chloroplast whole genome phylogeny also provides an overall comparison of the *Rosa* accessions that are currently available on Genbank and highlights specimens with a doubtful identification. All these results could now

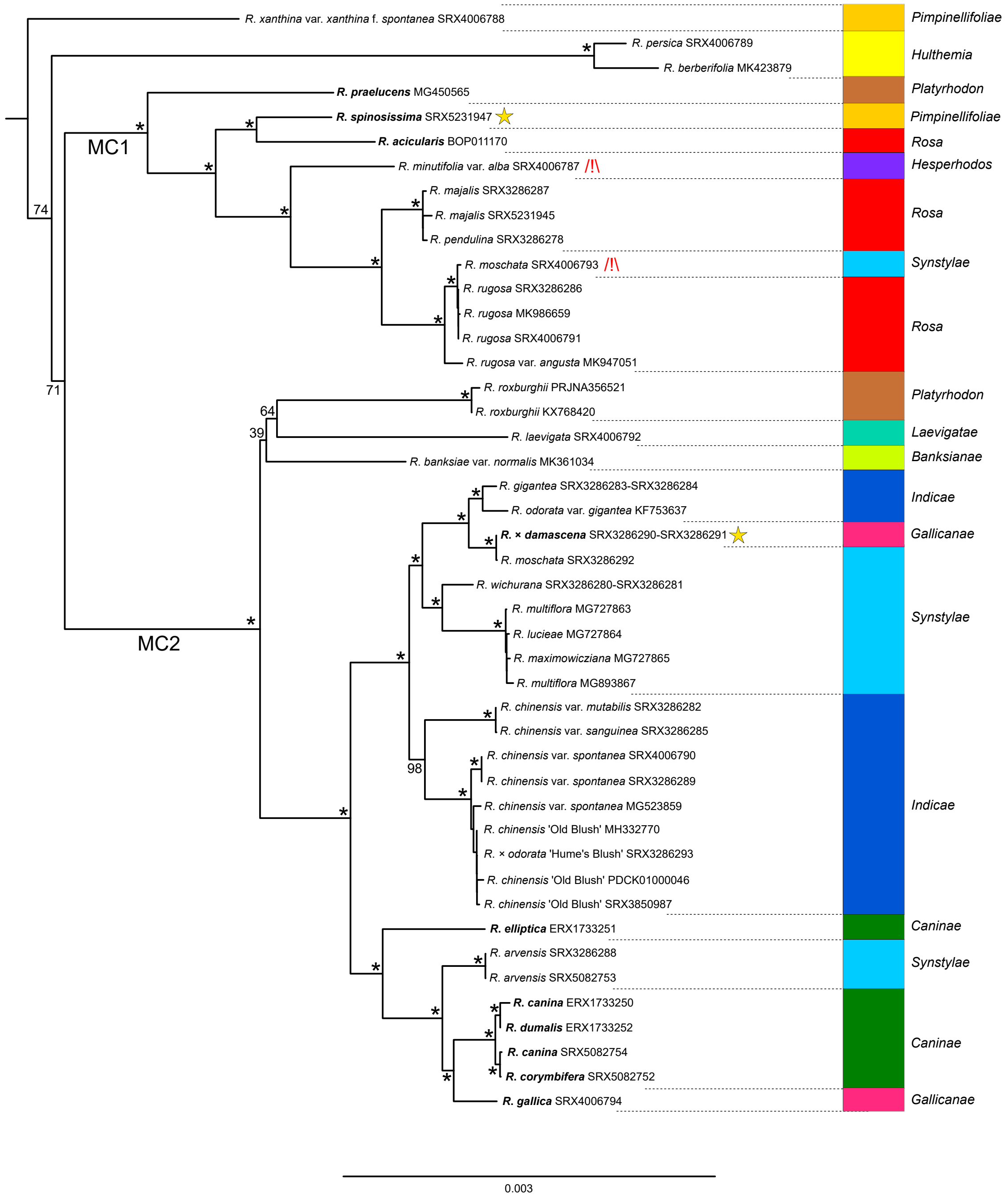

R. xanthina var. xanthina f. spontanea SRX4006788
R. persica SRX4006789
R. berberifolia MK423879
R. praelucens MG450565
R. spinosissima SRX5231947
R. acicularis BOP011170
R. minutifolia var. alba SRX4006787 /!\
R. majalis SRX3286287
R. majalis SRX5231945
R. pendulina SRX3286278
R. moschata SRX4006793 /!\
R. rugosa SRX3286286
R. rugosa MK986659
R. rugosa SRX4006791
R. rugosa var. angusta MK947051
R. roxburghii PRJNA356521
R. roxburghii KX768420
R. laevigata SRX4006792
R. banksiae var. normalis MK361034
R. gigantea SRX3286283-SRX3286284
R. odorata var. gigantea KF753637
R. × damascena SRX3286290-SRX3286291
R. moschata SRX3286292
R. wichurana SRX3286280-SRX3286281
R. multiflora MG727863
R. lucieae MG727864
R. maximowicziana MG727865
R. multiflora MG893867
R. chinensis var. mutabilis SRX3286282
R. chinensis var. sanguinea SRX3286285
R. chinensis var. spontanea SRX4006790
R. chinensis var. spontanea SRX3286289
R. chinensis var. spontanea MG523859
R. chinensis 'Old Blush' MH332770
R. × odorata 'Hume's Blush' SRX3286293
R. chinensis 'Old Blush' PDCK01000046
R. chinensis 'Old Blush' SRX3850987
R. elliptica ERX1733251
R. arvensis SRX3286288
R. arvensis SRX5082753
R. canina ERX1733250
R. dumalis ERX1733252
R. canina SRX5082754
R. corymbifera SRX5082752
R. gallica SRX4006794
Pimpinellifoliae
Hulthemia
Platyrhodon
Pimpinellifoliae
Rosa
Hesperhodos
Rosa
Synstylae
Rosa
Platyrhodon
Laevigatae
Banksianae
Indicae
Gallicanae
Synstylae
Indicae
Caninae
Synstylae
Caninae
Gallicanae
MC1
MC2
74
71
64
39
98
*
0.003

Figure 39: Phylogenetic tree obtained from Maximum Likelihood inference analysis of the chloroplast whole genome sequences of *Rosa*. Bootstrap Support (BS) (×1000) are presented as close as possible to their corresponding branch. * indicate a maximal BS of 100. Known polyploids from literature are in bold. Genbank ID is mentioned after each accession name. Yellow star indicates known intersectional hybrid from Chapter 3. /!\ indicates possible misidentified accessions. MC: meta clade.

be considered to sketch proposals to modify the current classification of the genus *Rosa*.

3.10.6 Additional file and availability of data and material

Additional File S1. Fasta alignment of the 45 chloroplast whole genome sequences of *Rosa* plus one outgroup (Fragaria vesca ssp. vesca) (Fasta file).

The aforementioned Additional File is available upon request to the GDO team.

3.11 Supplementary results Part B: Proposals for taxonomic modifications in the genus *Rosa* along with general considerations

In this part, we provide proposals for further taxonomic modifications in the genus *Rosa*, based on the recent phylogenetic analyses from Chapter 3 and the literature (Figure 40). For some sections, general discussions are developed to explain classification choices based on the evolutionary history of wild roses.

We suggest to divide the genus *Rosa* into sections based on the observations that subgenus *Rosa* in not monophyletic in every analysis and because other subgenera (*Hulthemia*, *Hesperhodos* and *Platyrhodon*) are embedded at different positions in the subgenus *Rosa*. Because of the strong morphological similarities between some currently recognized sections, we consider that treating them at the subgenus level will be a overestimation of distinctiveness. Therefore, we will hereafter treat subgenera *Hulthemia*, *Hesperhodos* and *Platyrhodon* as sections, a choice also suggested by Fougère-Danezan et al. (2015). Considering nomenclatural acts that may appear in this part, the different amendments to the denominations present in the text below should not be treated as formally validated and only correspond to suggestions that should be considered for a possible revision of the genus *Rosa*. Therefore, the names in the text below should be considered as "proposed in anticipation of the future acceptance of the taxon concerned" under Art 36.1 if International Code of Nomenclature for Algae, Fungi and Plants was applied (Turland et al., 2018).

For each taxa, we considered individually different aspects, namely change in its rank or fusion, change in its composition, infrasectional division, suggestion for name changes, putative hybridogenic origin of the section or of species, doubtful accessions in our study.

3.11.1 *R.* subg. *Rosa*

Change in taxonomic rank or fusion

The subgenus *Rosa* should no longer be recognized as we treat other subgenera as sections. The genus *Rosa* should only include sections based on phylogenetic and morphological arguments.

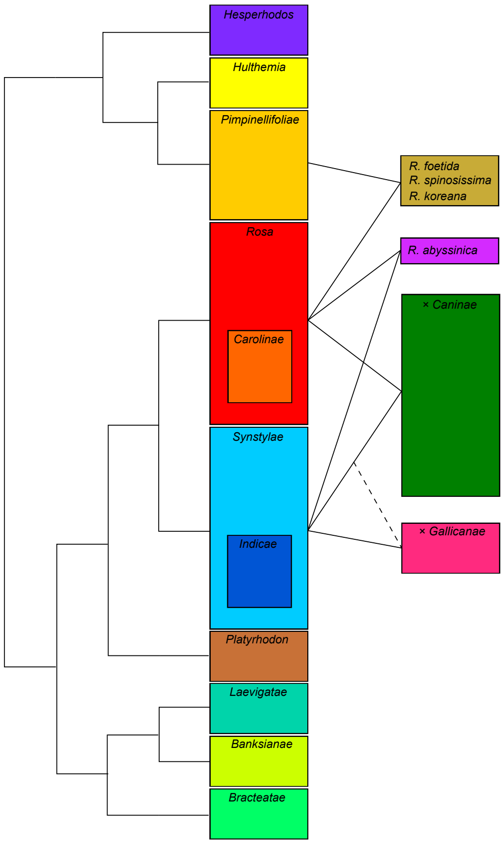


Figure 40: Schematic representation of the link between our nomenclature proposals and the observed phylogenetic relationships between *Rosa* taxa. Subgenera are treated as sections. The topology shows the relationships between core sections as presented in Chapter 3. On the right most part, hybrid sections or groups are presented along with their relationships with the core sections.

.

Change in composition

Should include all species of the genus *Rosa*.

Infrasectional division

Not applicable

Suggestion for name changes

None

Hybridogenesis

See sections

Doubtfull accessions

See sections

3.11.2 *R.* sect. *Banksianae*

Change in taxonomic rank or fusion

Should be maintained as a section.

Change in composition

None

Infrasectional division

Not needed

Suggestion for name changes

None

Hybridogenesis

Section not assumed to have a hybridogenic origin

Doubtfull accessions

R. banksiae (BAN04) is found closer to *Bracteatae* species than to the other accession of section *Banksianae* (*R. banksiae* BAN03).

3.11.3 *R.* sect. *Bracteatae*

Change in taxonomic rank or fusion

Should be maintained as a section.

Change in composition

None

Infrasectional division

Not needed

Suggestion for name changes

None

Hybridogenesis

Section not assumed to have a hybridogenic origin

Doubtfull accessions

R. bracteata (BRA03) grouped with *Caninae* accessions in both plastid and nuclear phylogenies and may correspond to rootstock material.

3.11.4 *R.* sect. *Caninae*

Change in taxonomic rank or fusion

Should be maintained as a section.

Change in composition

We showed that material assigned to *R. marginata* may have a hybrid origin (sect. *Rosa* × sect. *Caninae*) (Figure 38) so *R. marginata* may be excluded from section *Caninae*.

Infrasectional division

Plastid data: Our plastid analyses revealed two groups in the *Caninae* clade (C5a and C5b, Figure 34). C5a mainly encompasses species from subsections *Rubigineae* and *Vestitae* while C5b comprises subsections *Caninae*, *Trachyphyllae* and *Tomentellae*. We had no specimen from subsection *Rubrifoliae*, however, we could consider that they would have been grouped with subsections *Rubigineae* and *Vestitae* based on the plastid phylogeny presented in Fougère-Danezan et al. (2015). Fougère-Danezan et al. (2015) also identified two main *Caninae* clades in the plastid phylogeny: Ru (*Rubigineae*) and Ca (*Caninae*). The Ru clade approximately corresponds to our C5a clade whereas the Ca clade resembles our C5b clade. However, the repartition of the *Caninae* subsections between the two clades is sometimes different from one analysis to the other. Especially, both C5b and Ca are close to European *Synstylae* (*R. arvensis*, *R. sempervirens*) and some species of section *Gallicanae*, but they slightly differ in their content. The Ca clade only includes species from subsection *Caninae* while our C5b clade also encompasses *Trachyphyllae* and *Tomentellae* species. Wissemann and Ritz (2005) also obtained two *Caninae* clades in their plastid analysis. These two clades roughly separated subsections *Caninae*, *Tomentellae*, *Trachyphyllae* in one clade and subsections *Rubiginae*, *Rubrifoliae*, *Vestitae* in another clade, in line with our observations.

Nuclear data: Our phylogenetic analyses based on nuclear sequences (Figure 37) do not show a clear separation in two groups as the one observed in chloroplast-based studies. In our network analysis, several accessions are placed relatively apart (*Rosa mollis* (MOL01); *R. sherardii* (SHE01), *R. inodora* (INO01), *R.* 'Laxa' (LAX01)) but grouped together in clade C5a of the plastid phylogeny (Figure 34), except for *R.* 'Laxa' that is found in clade C5b. Other parts of the network also mix members of C5a and C5b (for example *Rosa orientalis* (ORI01) mainly linked with members of the plastid clade C5a; or *R. tomentosa* (TOA02) and *R. horrida* (HOR01) linked with members of the plastid clade C5b). Even if we consider putative miss-identification, for the same samples plastid and nuclear data do not provide the same relationships. Nuclear data performed poorly in further delineating section *Caninae*, in line with observations done by De Cock et al. (2008); De Riek et al. (2013). The authors analyzed the genetic structure of section *Caninae* with nuclear AFLPs and demonstrated that subsectional delineations in *Caninae* are difficult because subsectional boundaries are often blurred.

Summary: Based on the two-claded disposition often observed in chloroplast-based phylogenetic analyses, section *Caninae* could be separated in only two subsections. Subsections *Caninae*, *Tomentellae*, and *Trachyphyllae* could be collapsed in one subsection and subsections *Rubiginae*, *Rubrifoliae*, *Vestitae* could be merged in another subsection. However, interspecific crosses are very common in section *Caninae*, and a lot of intermediate specimens might be observed, some of which considered as hybrids between members of distinct subsections (*R. nitidula* = *R. canina* [*Caninae*] × *R. rubiginosa* [*Rubiginae*]; *R. scabriuscula* = *R. canina* [*Caninae*] × *R. tomentosa* [*Vestitae*]) (Bakker et al., 2019). Based on nuclear data, splitting the *Caninae* section in two subsections is less relevant. This could reveal a significantly different history between the nuclear and chloroplastic genomes or a bias from the pentaploid genomes and their particular *Caninae* meiosis.

Suggestion for name changes

None

Hybridogenesis

Nuclear analysis mostly revealed that section *Caninae* is likely the result of recent intersectional crosses between European species of sections Synstylae and *Rosa* (Zhang et al., 2013; Fougère-Danezan et al., 2015; Ballmer, 2018). In Chapter 3, we computed the respective contributions of sections *Rosa* and *Synstylae* to the formation of *R. agrestis* (clade C5a, subsect. Rubiginae) and *R. canina* (Clade C5b, subsect. *Caninae*). We concluded that the same ratios were inherited from sections *Rosa* and *Synstylae* in these two contrasted *Caninae* species. Our results suggested that section *Synstylae* contributed to 3/5 and section *Rosa* to 2/5 to the formation of the *Caninae* section, in line with the pentaploidy commonly observed in the section. Different hypothesis can be made regarding the origin of section *Caninae* and one is presented in Figure 41 . Our hypothesis assumes that two types of species have contributed to the formation of *Caninae*: species of type A (*R. arvensis*, European *Synstylae*) and species of type M (*R. majalis*, European *Rosa*). We favored a scenario that did not involve too high ploidy levels since hexaploidy remains rare in wild roses. The formation of unreduced gametes in a set of diploid species could have given rise to regular allotetraploid. As in (Crhak Khaitova et al., 2014), we also suggest the use of a triploid bridge to obtain a pentaploid specimen with ratios that correspond to those find in our hybrid networks. Our results do not support the statement that an extinct lineage (*Protocaninae*) would have contributed to the *Caninae* genome (Ritz et al., 2005; Crhak Khaitova et al., 2014). However, we acknowledge that many different intermediate forms (*R. jundzillii*, *R. glutinosa*) may occur and that the origin of section *Caninae* may be multiple as suggested in Ballmer (2018). These inheritance proportions should be verified for each dog rose species using a much larger number of accessions. We therefore encourage taxonomists to focus on the identification of sets of representative specimens for each of the six subsections described in Wissemann (2003a), and then resort to long read whole genome sequencing to identify homologous chromosome regions between representative specimens that would further help understand the evolutionary history of dog roses.

Doubtfull accessions

R. montana (MON01) is embedded within clades of polyploid species from section *Rosa*. In addition, we found *R. montana* (MON01) to be tetraploid which is less common than pentaploidy in *Caninae* section. *R. montana* (MON01) may therefore be misidentified.

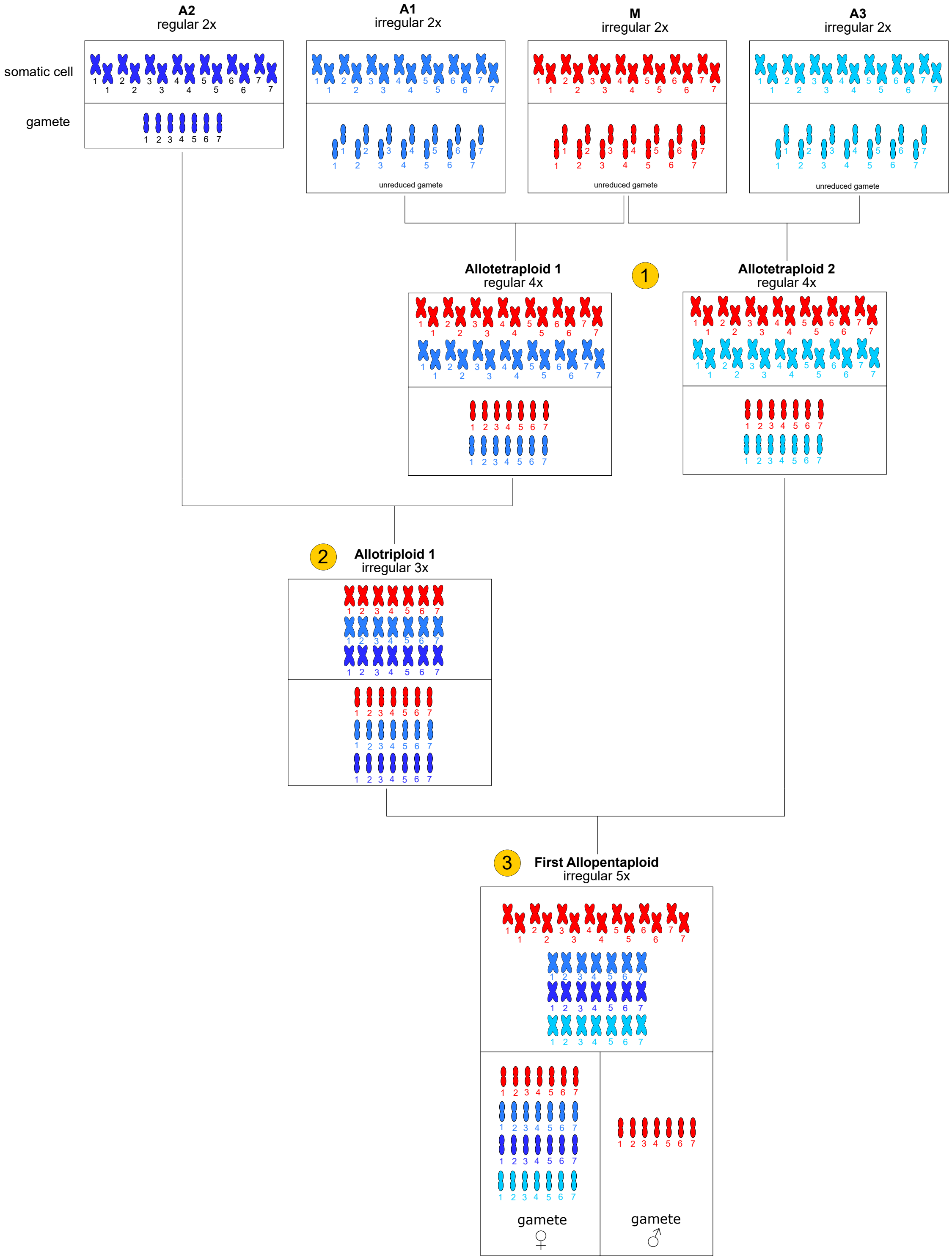
3.11.5 *R.* sect. *Carolinae*

Change in taxonomic rank or fusion

All *Carolinae* species are embedded within a clade of North American species from section *Rosa* in several analyses (Bruneau et al., 2007; Fougère-Danezan et al., 2015) as well as in ours. Therefore, on the base of a purely phylogenetic argument, we suggest to merge section *Carolinae* with section *Rosa* as it has already been suggested before (Fougère-Danezan et al., 2015), even if morphologically there exist slight differences between the two sections (Joly and Bruneau, 2007).

Change in composition

None



(Caption next page.)

Figure 41: (Previous page.) Possible original hybridization processes behind section *Caninae* following results from Chapter 3. The scenario involves two sections: *Rosa* (red, one species: M) and *Synstylae* (blue, 3 species or forms: A1, A2, and A3) and consists of three steps. Step 1: Diploid species in both sections form unreduced gametes that after fertilization give rise to allotetraploid individuals. The subgenomes from both parents juxtapose but do not mix because they are too heterogeneous. Step 2: One of the allotetraploids crosses with a regular diploid of the *Synstylae* section forming an allotriploid. Step 3: Allotriploid produces unreduced gametes and crosses with an allotetraploid from step 1. At this stage, the bivalence of the red chromosomes is restored while the three blue chromosome sets remain univalent due to their too large difference. Only bivalent chromosomes are mixed during meiosis and univalent chromosomes do not segregate. These are transmitted by the maternal lineage. The presented scenario displays a two-fifth genome proportion that is inherited from section *Rosa* and a three-fifth genome proportion that is inherited from section *Synstylae*.

Infrasectional division

Not needed

Suggestion for name changes

None

Hybridogenesis

Hybridizations spontaneously occur between North American species from section *Rosa* and species from section *Carolinae* (Joly et al., 2006b). *R. carolina* may have a reticulate origin (Joly et al., 2006b) and we found that *R. palustris* can may also have a hybrid origin. In addition, we observed that *R. carolina* may have greatly contributed to *R. californica* (from section *Rosa*).

Doubtfull accessions

None

3.11.6 *R.* sect. *Chinenses* (syn. *Indicae*)

Change in taxonomic rank or fusion

Section *Chinenses* is often found embedded in section *Synstylae* in plastid based analysis (Fougère-Danezan et al., 2015; Zhu et al., 2015). Here, we demonstrated that it is also the case with nuclear sequences. Therefore, section *Chinenses* should be merged with section *Synstylae*. If the name *Synstylae* is anterior to *Chinenses* then *Synstylae* could be selected as the name of this merged group.

Change in composition

None

Infrasectional divisions

Not needed

Suggestion for name changes

None

Hybridogenesis

Section *Chinenses* includes two species: *R. chinensis* and *R. odorata*. However, each of these two species have numerous varieties that were extensively used to create cultivars through hybridization. The wild origin of each variety has to be assessed.

Doubtfull accessions

None

3.11.7 *R.* sect. *Gallicanae*

Change in taxonomic rank or fusion

None

Change in composition

None

Infrasectional divisions

Given the hybrid origin of this section, future studies may further delineate groups in section *Gallicanae* based on the contribution of the other sections to each *Gallicanae* species.

Suggestion for name changes

None

Hybridogenesis

Section *Gallicanae* encompasses species which wild origin can be debated. *R. gallica* probably thrived in Europe after that some cultivated specimens escaped from the Romans' rose gardens during Antiquity. *R. gallica* seems to be a 50/50 hybrid between a European lineage of section *Synstylae* and a common ancestor to *Synstylae* and *Caninae* that might not exist anymore. Other *Gallicanae* species including *R.* × *damascena* and *R.* × *alba* likely correspond to non-natural hybrids obtained after crossing *R. gallica* with species from other sections (*Synstylae* (Iwata et al., 2000) and probably *Caninae* (Zlesak, 2009)). More work is needed, especially on the origin of *R. gallica* to assess its wild origin.

Doubtfull accessions

None

3.11.8 *R.* subg. *Hesperhodos*

Change in taxonomic rank or fusion

From a phylogenetic perspective, the subgenus *Hesperhodos* should be treated as a section since it is embedded in the subgenus *Rosa*.

Change in composition

None

Infrasectional divisions

Not needed

Suggestion for name changes

The new section should be best named *Hesperhodos*.

Hybridogenesis

None

Doubtfull accessions

The existence of wild *R. minutifolia* var. *alba* has to be verified.

3.11.9 *R.* subg. *Hulthemia*

Change in taxonomic rank or fusion

From a phylogenetic perspective, the subgenus *Hulthemia* should be treated as a section since it is embedded in the subgenus *Rosa*. Moreover, the existence of wild hybrids between *R. persica* and section *Rosa* has been demonstrated (Vaezi et al., 2019) thus reinforcing the idea that subgenus *Hulthemia* is not so distant from subgenus *Rosa*.

Change in composition

None

Infrasectional divisions

Not needed

Suggestion for name changes

The new section should be best named *Hulthemia*.

Hybridogenesis

None

Doubtfull accessions

None

3.11.10 *R.* sect. *Laevigatae*

Change in taxonomic rank or fusion

Should be maintained as a section

Change in composition

None

Infrasectional divisions

Not needed

Suggestion for name changes

None

Hybridogenesis

Section not assumed to have a hybridogenic origin

Doubtfull accessions

None

3.11.11 *R.* sect. *Pimpinellifoliae*

Change in taxonomic rank or fusion

The core *Pimpinellifoliae* clade should be treated as a section.

Change in composition

The polyphyly of section *Pimpinellifoliae* has been reported in different phylogenetic analysis based on both nuclear and plastid data (Matsumoto et al., 2000b; Wissemann and Ritz, 2005; Bruneau et al., 2007; Koopman et al., 2008; Fougère-Danezan et al., 2015). At least two species (*R. spinosissima* and *R. foetida*) often branch elsewhere in the *Rosa* phylogeny, away from a core *Pimpinellifoliae* clade (Bruneau et al., 2007; Koopman et al., 2008; Fougère-Danezan et al., 2015). The core *Pimpinellifoliae* clade generally comprises diploid species from the *R. sericea* complex (*R. sericea*, *R. omeiensis*, *R. zhongdianensis*, *R. sikangensis*, *R. morrisonensis*, *R. taronensis*, *R. mairei*) and yellow-flowered diploid species such as *R. xanthina*, *R. hugonis*, *R. primula*, *R. ecae* (Jan et al., 1999; Qiu et al., 2012; Fougère-Danezan et al., 2015). *R. tsinglingensis* has been reported sister to species from section *Rosa* in both plastid and nuclear phylogenies (Fougère-Danezan et al., 2015) and might be misplaced in the *Pimpinellifoliae* section, as for *R. farreri* (Roberts, 1977). More work should be carried out on *R. ecae*, *R. hemisphaerica* var. *rapinii*, *R. koreana*, and *R. kokanica* to assess their correct assignment to the core *Pimpinellifoliae*. Especially, *R. koreana* that has been found sister to species from section *Rosa* in Bruneau et al. (2007); Fougère-Danezan et al. (2015) with different accessions. *R. koreana* may therefore be a hybrid between the core *Pimpinellifoliae* clade and section *Rosa*. The same conclusion may be developed for *R. kokanica*, that was found near section *Rosa* in plastid analysis and near the core *Pimpinellifoliae* group in nuclear analysis. However, to the best of our knowledge, our study was the first to include *R. kokanica* and our observations need to be confirmed by sampling additional specimens of this species.

Infrasectional divisions

Future taxonomic modifications should separate *R. spinosissima* and *R. foetida* from the core *Pimpinellifoliae* clade since we demonstrated that they likely correspond to intersectional hybrids between sections *Rosa* and *Pimpinellifoliae*. Section *Pimpinellifoliae* should therefore only include species from the *R. sericea* complex, along with *R. primula*, *R. xanthina* and *R. hugonis*.

Suggestion for name changes

Keeping the name *Pimpinellifoliae* for the core *Pimpinellifoliae* would not be possible since *R. spinosissima* (syn. *R. pimpinellifolia*) is the type species of this section. If a new section name is attributed for species from the core *Pimpinellifoliae* clade, section *Pimpinellifoliae* could therefore only include species that share intermediate characters between section *Rosa* and the core *Pimpinellifoliae* clade, such as *R. spinosissima*.

Hybridogenesis

Polyploid *Pimpinellifoliae* species are generally found outside the core *Pimpinellifoliae* in plastid phylogenies, often embedded in a clade that roughly corresponds to section *Rosa* (Matsumoto et al., 1998; Wissemann and Ritz, 2005; Bruneau et al., 2007; Fougère-Danezan et al., 2015). The proximity between sections *Pimpinellifoliae* and *Rosa* has also been reported in nuclear phylogenies (Wu et al., 2001; Scariot et al., 2006; Meng et al., 2011; Zhu et al., 2015), and here in hybrid networks (Chapter 3). Sections *Pimpinellifoliae* and *Rosa* may therefore have good affinity to spontaneously hybridize. Taking the example of *R. spinosissima*, this species has the northernmost area of distribution of all *Pimpinellifoliae* and overlap those of northern *Rosa* species (*R. acicularis*, *R. majalis*). The polyphyly of section *Pimpinellifoliae* may be the result of recurrent intersectional crosses, preferentially with northern species of the section *Rosa*. Considering *R. spinosissima* and its varieties (nine *sensu* Masure (2013), both plastid and nuclear phylogenetic analyses of Wissemann and Ritz (2005) showed that *R. altaica* (syn. *R. spinosissima* var. *altaica*) and *R. spinosissima* do not form a monophyletic group, each of these two taxa being close to different diploid species from section *Rosa*. According to Wissemann and Ritz (2005), although these two taxa can currently be separated morphologically only by size, they are genetically distinct (both species are tetraploid, at least *R. spinosissima* seems to be allotetraploid. For *R. altaica*, Wissemann and Ritz (2005) report putative cryptic hybridization between a *R. spinosissima*-derivate and a member of section *Cinnamomeae* [i.e. sect. *Rosa*]. The large number of natural *R.* spinosissima varieties (nine referenced in Masure (2013)) suggests that this species may have spontaneously appeared at different place and time in the northern hemisphere from crosses between lineages of sections *Rosa* and *Pimpinellifoliae*. The resulting "varieties" could therefore display substantial variations as reported between *R. spinosissima* (Coastlines and British Isles) and *R. spinosissima* var. *altaica* (Altai mountains). The putative polytopic origin of *R. spinossima* morphotypes would be interesting to investigate in terms of genomic structure and ecology. From plastid phylogenies, the maternal lineage of *R. spinosissima* seems to originate from section *Rosa* and thus the pollen parent would be from the core *Pimpinellifoliae* clade.

Doubtfull accessions

R. hemisphaerica var. *rapinii* (HES01) is embedded in *Synstylae* clades in both nuclear and plastid analyses and may therefore be misidentified. *R. koreana* (KOR02) likely corresponds to rootstock material from section *Caninae*.

3.11.12 *R.* subg. *Platyrhodon*

Change in taxonomic rank or fusion

Subgenus *Platyrhodon* has always been found embedded with subgenus *Rosa* (Fougère-Danezan et al., 2015) and should therefore be treated as a section.

Change in composition

It should exclude the decaploid *R. praelucens* since it does not cluster with diploid *R. roxburghii* in either plastid or nuclear analysis. Moreover, we demonstrated that *R. praelucens* is likely an intrasectional hybrid between different species from section *Rosa* with no intervention of *R. roxburghii* (Figure 38). Therefore, subgenus *Platyrhodon* should only include *R. roxburghii* and its natural varieties.

Infrasectional divisions

Not needed

Suggestion for name changes

None

Hybridogenesis

None

Doubtfull accessions

None

3.11.13 *R.* sect. *Rosa*

Change in taxonomic rank or fusion

None

Change in composition

Should include section Carolinae, and the species *R. praelucens* and *R. farreri.* Other species assignments remain unchanged.

Infrasectional divisions

We delimited three clades in our plastid analysis (Chapter 3): C2a, C2b and C2c. C2a and C2b largely correspond to polyploid clades with mostly Asian species while C2c comprises a balanced ratio between diploid and polyploid species from North America (*Carolinae* fully included in this clade). In the nuclear analysis, we also observed a segregation between Asian and North American species from section *Rosa*, however we do not see the value in further dividing the *Rosa* section into subsections. C2a, C2b and C2c form a well-supported meta clade that encompasses nearly all species from section *Rosa*, except some doubtful accessions.

Suggestion for name changes

None

Hybridogenesis

Section *Rosa* is the largest section of the genus and encompasses diploid and a wide range of polyploid species with different ploidy levels. Some clades nearly encompass only polyploid taxa (C2b). Given the ability of wild roses to spontaneously hybridize, we could suppose that most polyploid species in section *Rosa* have an allopolyploid origin. However, we developed few results on the hybrid origin of polyploid taxa in section *Rosa* and more work should be carried on to investigate reticulate patterns within this section. Extensive diploid interspecific crosses have also been reported between *R. gymnocarpa* and other members of section *Rosa* (Ertter and Lewis, 2016).

Doubtfull accessions

Several accessions originally attributed to section *Rosa* can be considered misidentified or originating from rootstocks (*R. giraldii* (GIR02) and *R. rugosa* (RUG02) within *Synstylae*; *R. bella* (BEL01) within *Caninae*; *R. moyesii* (MOY03) and *R. hemsleyana* (HEM02) with the core *Pimpinellifoliae* clade).

3.11.14 *R.* sect. *Synstylae*

Change in taxonomic rank or fusion

None

Change in composition

Should include section *Chinenses* (syn. *Indicae*). *R. abyssinica* may be excluded from section *Synstylae* since it likely correspond to an intersectional hybrid (sect. *Rosa* × sect. *Synstylae*), as shown in cross comparisons between plastid and nuclear phylogenies in Chapter 3 and other studies (Fougère-Danezan et al., 2015; Zhu et al., 2015).

Infrasectional divisions

Almost all species in section *Synstylae* (including Indicae species) are diploid and their phylogenetic relationships have been widely studied in Zhu et al. (2015). The authors discerned no less than 10 different groups (*Indicae*, *Rosa abyssinica,* North American *Synstylae*, European *Synstylae*, *R. glomerata*, *R. rubus* group, *R. soulieana* group, *R. helenae* group, *Rosa multiflora* group, and *Rosa longicuspis* group). We disagree with (Lewis, 2016) about the placement of *R. setigera* in a new section *Americanae*. Our nuclear phylogenetic analyses show that *R. setigera* is largely a member of the *Synstylae* section, although we acknowledge that it branches at the basis of the core *Synstylae* group which is in line with the results from Zhu et al. (2015).

Suggestion for name changes

If a further division of section *Syntylae* is desired, we therefore advocate to split the section in three subsections that may be named: “Americanae” (*R. setigera*), “Europae” (*R. arvensis*, *R. sempervirens*, *R. leschenaultiana*), and “Asiae” (Asian *Synstylae* and *Indicae*) as an indication of the geographic distribution of their members.

Hybridogenesis

Intrasectional hybridizations are quite common between *Synstylae* species (*R. glomerata* (GLO01), *R. moschata* (MOS02), *R. soulieana* (SOU02) reported in Chapter 3, *R. glomerata*, *R. lichiangensis*, *R. multiflora* × *R. rubus*, *R. rubus* × *R. lucidissima*, reported in Zhu et al. (2015)). Moreover, *Synstylae* species seem to maintain a high level of heterozygosity as suggested in the distributions of allele frequencies of *R. filipes* (FIL01), *R. giraldii* (GIR01, GIR02), *R. moschata* (MOS02), *R. mulliganii* (MUI01), *R. rubus* (RUB01), and *R. soulieana* (SOU01).

Doubtfull accessions

R. anemonaeflora (ANE01) and *R. glomerata* (GLO01) may be misplaced within species from section *Rosa* due to either misidentification or hybrid origin.

4 Close species relationships in *Rosa*: what can we learn from microsatellite (SSRs) markers?

4.1 Preamble

Chapter 3 provided a general framework to obtain robust phylogenetic hypotheses while considering evolutionary processes such as hybridization and polyploidy within the large and complex genus *Rosa*. Although phylogenetic relationships between sections are now well understood, many closer relationships at the infrasectional level remain unclear. I hypothesized that phylogenetics can no longer bring substantial information at such shallow taxonomic ranks. Indeed, close species in the phylogenetic tree are likely to share similar distributions, habitats, and characteristics. The reproductive barriers between such taxa are supposed to be very loose, with a substantial amount of gene flows. Moreover, incomplete lineage sorting may be prominent at such recent times. Therefore, the phylogenetic frame may mismodel recent speciations and additional analyses must be carried to better characterize the relationships between closely-related species.

During the sampling steps for Chapter 3, I struggled with obtaining wild material of *Rosa persica* since this species hardly survive in American-European botanic gardens within easy reach. I therefore contacted several researchers in Iran that previously worked on the genetic diversity of *Rosa persica* and asked for few quality samples. This is basically how I met Dr Zahra Karimian, assistant professor at the Ferdowsi University of Mashhad. As she sent me excellent material from this species, I came up with the idea to further extend our collaboration in the frame of a project. At that time, the University of Angers called for project proposals to strengthen international collaborations. I decided to write a small project that would aim at studying the genetic diversity of *Rosa persica* with the idea to unveil the origin of this species, bringing two hypotheses into opposition: *Rosa persica* is a living fossil of the genus *Rosa* vs *Rosa persica* is a super-evolved species, well adapted to its habitat. I received a € 5.5k grant from the University to investigate this subject, allowing me to invite Zahra for one month in France in October 2018. She brought several samples of *Rosa persica* from diverse regions in Iran. Given the time and money availabilities, I suggested that genotyping the *Rosa persica* accessions together with species with an overlapping distribution would be the best experiment that we could do. Section *Pimpinellifoliae* has numerous species that share the same habitats as *R. persica*, therefore I decided to genotype almost all of the *Pimpinellifoliae* accessions present in my collection. Concerning the genotyping, I used the same set of SSRs as Liorzou et al. (2016) did on cultivated material. I imagined that it would also be interesting to compare the alleles she found in cultivated roses with those we would discover in their wild counterparts. After Zahra came, I hired a bachelor trainee, Éléonore Malé, for two months to help with the genotyping and analyses.

The purpose of this chapter is to highlight the work produced during this small project, which was carried out somewhat apart from my thesis. I decided to write a small article (this chapter) that would allow (1) to assess the resolving power of SSRs to delimit complexes of closely related species by considering three proximity scales: the genus (*Rosa*), the section (*Pimpinellifoliae-Hulthemia*) and the species (*Rosa persica*) and (2) would allow the GDO team to easily compare the wild alleles with those identified by Liorzou et al. (2016) in the cultivated pool.

The authors who contributed to this work are:

Kevin Debray[1], Éléonore Malé[1], Zahra Karimian[2], Annie Chastellier[1], Fabrice Foucher[1] and Valéry Malécot[1].

[1] IRHS, Agrocampus-Ouest, INRA, UNIV Angers, SFR 4207 QuaSaV, Beaucouzé, France

[2] Department of Ornamental Plants, Research Center for Plant Sciences, Ferdowsi University of Mashhad, Iran

The contribution of each author is detailed at the end of this chapter.

4.2 Introduction

Phylogenetic relationships between closely related taxa are often difficult to unveil, especially at the infrageneric level. There are two main aspects that limit the use of a phylogenetic framework to study relationships between sister species that diverged recently (<5MYa): (1) limitations related to the methods and (2) limitations related to the biology of the taxonomic group. Sister species are likely to have diverged recently and may still share a large number of common features that hamper a clear distinction between taxa. Some evolutionary processes, such as hybridization and incomplete lineage sorting (ILS), result in non-tree like patterns (Gallardo, 2017). Such reticulate patterns are barely considered in molecular studies of phylogenetic relationships, yet hybridization may concern a substantial number of closely related species (Mallet, 2005; Abbott et al., 2013), especially in plants (Soltis and Soltis, 2009; Wissemann, 2010; Alix et al., 2017). Instead, evolutionary histories involving reticulations are fitted into an inappropriate bifurcating tree, thus resulting in poorly supported branches at recent times. Phylogenetic studies at the Sanger sequencing era traditionally involve few molecular sequences that generally show little variations between closely related taxa. A clear delineation of species boundaries is therefore limited. Finally, phylogenetic reconstructions involve the choice of an evolutionary model of molecular sequences. However, classic evolutionary models might not be optimized to study recent speciation events since they generally assume that all sites in the alignment have evolved along the same phylogenetic tree. Therefore, branches of closely related taxa may show less support than deeper branches in the phylogenetic tree due to the limitations of DNA evolution models. The concatenation of large arrays of gene sequences through phylogenomics (Bleidorn, 2017; Patané et al., 2018) may apparently enable to obtain stronger relationships for closely related taxa. However, a thorough analysis of each gene tree individually generally show an extensive amount of conflicts (Jeffroy et al., 2006; Smith et al., 2015), especially toward recent speciations (Debray et al., 2019). Therefore, resorting to phylogenetic/phylogenomic studies might not be the most effective strategy to recover a comprehensive insight into very close species relationships. For populations of individuals belonging to a same species, several methods already exist to study the genetic diversity and structure of the group (Charlesworth and Charlesworth, 2017; Casillas and Barbadilla, 2017; Grünwald et al., 2017; Tomasello, 2018). These methods are commonly used in population genetics and provide a detailed perspective of the possible gene flows between subgroups of individuals. They are mostly based on distance matrixes derived from the sharing of same variations of genotypic markers. Genotyping still represents a rapid and cost-effective strategy to access genetic diversity for a large number of accessions. There exists several kinds of molecular markers (Nadeem et al., 2017) among which Single Sequence Repeats (SSRs or microsatellites) present numerous advantages such as (1) the requirement of a small quantity of genomic DNA for SSR detection, (2) high polymorphism, meaning that just tens of SSRs possibly enable to discriminate very close individuals, (3) a wide distribution in the genome and (4) a detection that can be automated with the possibility to multiplex (Rahman et al., 2009; Vieira et al., 2016). For an equivalent workload and cost, genotyping a few SSRs may be more effective than sequencing a few gene fragments to recover a large number of polymorphisms. Therefore, SSRs seem to be an ideal solution to distinguish closely related individuals, and already proved to be useful in plant barcoding (Chinnappareddy et al., 2012; Li et al., 2018). However, contrary to phylogenetic methods, distance-based population studies with molecular markers generally do not include evolutionary models. Therefore, distance-based studies convey phenetic relationships, ie based on a degree of similarity, rather than evolutionary relationships, ie based on the state of characters at each sequence position (Sokal, 1986; de Queiroz and Good, 1997).

Studying the phylogenetic relationships of species belonging to the genus *Rosa* is difficult because there exist about 150-200 wild rose species that easily hybridize (Rehder, 1940; Wissemann, 2003a). Recent analyses provided new insights into the phylogenetic relationships between sections and subgenera, enabling to highlight many patterns of intersectional hybridizations (Fougère-Danezan et al. (2015), Chapter 3). However, clear species relationships at the intrasectional level still remain challenging in *Rosa* phylogenetics. Genotyping studies already proved to be useful in some areas of the *Rosa* phylogeny, especially at the intrasectional level. Species delineation in the *Caninae* complex has been investigated using AFLP markers and enabled to distinguish genetically distinct groups, although the definition of a clear boundary for each group was challenging (De Cock et al., 2008; De Riek et al., 2013). Genetic relationships were also investigated in the complex of Alpine shrubs (*R. sericea*, *R. omeiensis*, *R. zhongdianensis*, *R. sikangensis*, *R. morrisonensis*, *R. taronensis*, *R. mairei*) using nuclear SSRs (Gao et al., 2015, 2019). In combination with molecular sequences and ecological aspects, the authors suggest that species belonging to the complex should be considered as one species, namely *R. sericea.* In all studies, the authors suggest that the transition from one species to another is not straightforward but rather corresponds to a continuum of intermediate genotypes. SSRs were also used to assess the genetic diversity of cultivated roses and enabled to highlight a shift over time from a European to an Asian genetic background (Liorzou et al., 2016).

Section *Pimpinellifoliae* encompasses diverse rose species from Asia to Europe, through Anatolia. Most phylogenetic studies on the genus *Rosa* report the polyphyly of section *Pimpinellifoliae* (Matsumoto et al., 2000b; Bruneau et al., 2007; Fougère-Danezan et al., 2015) which might indicates the presence of contrasted subgroups that would be interesting to investigate using genotyping methods. A recent analysis on nuclear sequences showed that a group of Asian *Pimpinellifoliae* is at the base of the genus *Rosa*, along with subgenera *Hesperhodos* and *Hulthemia* (Chapter 3). The former is native to North America (Baja California) and contains two species (Rehder, 1940; Lewis, 1965) while the latter originates from the Middle East and corresponds to one species: *R. persica* (Rehder, 1940). Both subg. *Hesperhodos* and *Hulthemia* comprise arid-adapted species, however, nuclear analyses showed that subg. *Hulthemia* is closer to the Asian *Pimpinellifoliae* than to the other arid-adapted subg. *Hesperhodos* (Chapter 3). The proximity between sect. *Pimpinellifoliae* and subg. *Hulthemia* is also reflected through other features such as (1) adjacent geographical areas with some overlapping regions in central Asia, (2) the sharing of species with bright-yellowed flowers, which color is found nowhere else in the genus.

We therefore assess the genetic diversity of *Rosa* species with emphasis on sect. *Pimpinellifoliae* and subg. *Hulthemia* using 32 microsatellite markers (SSRs). Our aims were (1) to investigate the extent to which SSRs can delineate different taxonomic ranks from the genus to the species levels and (2) to study possible gene flows between subg. *Hulthemia* and sect. *Pimpinellifoliae.*

4.3 Materials and Methods

4.3.1 Plant material and nomenclature

A total of 91 accessions representing 52 different species and varieties of the genus *Rosa* were sampled for our study (Appendix C, Supplementary Table C.1). Most of the samples are from section *Pimpinellifoliae* and subgenus *Hulthemia*, however, almost all subgenera and sections of *Rosa* were represented through at least one species. Accessions of *R. persica* (*R.* subg. *Hulthemia*) were mainly collected in different localities of Iran. The remainder accessions were either collected in the wild or provided by botanic gardens. We favored the selection of accessions with a known wild origin. Accessions that do not belong to either section *Pimpinellifoliae* or subgenus *Hulthemia* were selected on the basis of their known or supposed geographical origin, favoring a geographical area from southern Europe to western China, through the Middle East. Leaf fragments were collected and immediately silica-dried to preserve DNA structure. Plant material are stored at the IRHS collection, Angers, France (see Chapter 3). The nomenclature used corresponds to the one proposed by Masure (2013) and that largely relies on Brumme et al. (2013). Morphological and biogeographical data were also collected from the literature (Roberts, 1977; Wissemann, 2003a; Basaki et al., 2009; Masure, 2013; Singh et al., 2017) for comparison with genotyping analyses. We selected nine qualitative traits with different categories: shrub architecture, maximum altitude, maximum shrub height, type of inflorescence, flower color, prickle shape, stipule attachment, mean number of leaflets, and hip form.

4.3.2 DNA extraction and genotyping

DNA extractions and validations followed the methods proposed in Chapter 3. We used the 32 SSRs developed in Liorzou et al. (2016) to investigate the genetic diversity among our wild rose accessions (except that we replaced RMS140 by H17C12 (Genbank ID: EC586469, developed in Gar et al. (2011))). The 32 SSRs are well distributed over the genome, each of the seven chromosomes (linkage groups) being covered by 4-6 SSRs. Each SSR is associated to a primer pair enabling the amplification of the SSR locus using PCR. SSRs were amplified using a 4-plex PCR, meaning that four SSR loci are amplified at the same time in each DNA well. Each primer pair is distinguished by a different fluorescent label. The final reactional mix per well has a volume of 5 µL and contains 2,5 µL of 1X Qiagen® Multiplex PCR Master Mix, 0.5 µL of the respective primer pairs mix (each primer sequence being at 2 µM) and 2 µL of DNA diluted at the optimal concentration (see methods in Chapter 3). The following PCR program was implemented in thermal cyclers (BIO-RAD S1000TM and C1000TM Thermal Cycler): 1m30s of preliminary denaturation at 95°C, followed by a series of 30s at 95°C, 1m30s at 55°C (ramping from 95°C to 55°C by 1°C.s^{-1}) and 75s at 72°C (ramping from 55°C to 72°C by 1°C.s^{-1}) repeated 35 times, then a final elongation of 1m30s at 72°C. Each 96-well PCR plate contained 5 controls (4 positive controls: *R.* 'Black Baccara', *R.* 'Old Blush', *R.* × wichurana, *R.* 'The Fairy', and one negative control: water). PCR products were analyzed using an ABI 96-capillary 3730XL DNA Analyser (ABI Prism, Applied Biosystems, Foster City, CA, USA) at the Gentyane platform, Clermont-Ferrand, France. Allele scoring was performed in two times using GeneMapperTM v.4.1 (Applied Biosystems®) with a first automated analysis and then a thorough manual curation to eliminate artefacts. The automated analysis relied on the bin set created for cultivated material in Liorzou et al. (2016). New alleles were expected since our study focuses mainly on wild material. This is why the visual curation enabled to (1) create new bins for new alleles, (2) adjust the size of the bin for alleles occurring at the boundaries of existent bins, (3) eliminate artefacts inherent to SSR multiplexing (ie when one of the four SSR displays a high fluorescence peak on its corresponding electropherogram, a shadow may be thrown onto the electropherograms of the 3 remaining SSRs. In this case, the automated process may detect a SSR allele while it actually corresponds to an artefact). It was difficult to estimate the SSR allele dosage, especially in polyploid taxa, therefore alleles were coded as presence/absence using custom script. A SSR was considered successful if at least 50% of the sampling could recovered at least one allele at the SSR locus. SSRs that do not match the above criteria were discarded.

4.3.3 Diversity analysis

The genetic diversity of the samples was assessed by calculating the number of observed alleles (Ao), the mean number of alleles per individual (Am), the effective number of alleles (Ae), and the number of rare alleles for each SSR. The following formula was used to calculate Am for SSR α:

$$Am_{\alpha} = \frac{\sum n_{k\alpha}}{N_{\alpha}}$$

where $n_{k\alpha}$ is the number of alleles carried by individual k for SSR α and N_{α} is the number of genotyped individuals for SSR α.

Ae for SSR α was then calculated according to the formula from Hamrick and Godt (1990):

$$Ae_{\alpha} = \frac{1}{\sum(\frac{N_{i\alpha}}{N_{\alpha}})^2}$$

where $N_{i\alpha}$ is the number of individuals carrying allele i for SSR α and N_{α} is the number of genotyped individuals for SSR α. Alleles were considered rare when they were present in <5% of the individuals. We also computed the number of unique alleles, meaning alleles that appeared only once and which are therefore not informative.

4.3.4 Distance analysis

Genetic distance

The Dice method was used to measure the genetic distance between our set of diploid and polyploid accessions and proved to be useful in Liorzou et al. (2016). The distances were calculated using 1000 bootstraps with the DARwin v6 software package (Perrier and Jacquemoud-Collet, 2006). The input corresponded to the allele presence/absence matrix. Distances between accessions were represented using non-metric multidimensional scaling (NMDS) in R using the package vegan with two dimensions and maximum 1000 iterations to identify a convergent solution (Oksanen et al., 2018). The stress value was used as a proxy for assessing the goodness-of-fit of the NMDS.

Morphological and biogeographical distance

We discretized morphological and biogeographical data into categories that were further converted in a matrix of presence/absence. The presence/absence matrix was uploaded in Darwin v6 to compute the Dice's distance using 1000 bootstraps. The resulting Dice distance matrix was used for comparison with Dice genetic distance and to cluster accessions based on their morphological distances. We used the k-medoids method with the R package cluster and the function *pam* to identify clusters of individuals based on their morphological distance (Maechler et al., 2018). We chose the optimal number of clusters based on the silhouette method using the function *fviz_nbclust* in the R package factoextra (Alboukadel and Mundt, 2017). The Principal Coordinate Analysis resulting from the k-medoid clustering was plotted to visualize the morphological distance between accessions.

Spatial distance

Precise spatial coordinates of most samples of *R. persica* were available. We computed pairwise spatial distance between each sample using custom script. The spatial distance corresponds to the minimal kilometer distance that separates two samples. Only the accession of PER19 had a vague localization (Afghanistan) so we arbitrarily selected a spatial coordinate in the middle of the country. Spatial coordinates were reported onto a map drawn in QGIS v3.8.3 and based on Natural Earth datasets (`https://www.naturalearthdata.com`). The map was hill-shaded to show the reliefs and possible geographical fences.

4.3.5 Structure analysis

The allele table from GeneMapper was modified using the R package polysat (Clark and Jasieniuk, 2011) with maximum 10 alleles per SSR to meet STRUCTURE's requirements for file input. The resulting allele file was used in STRUCTURE v2.3.4 to investigate population structure (Pritchard et al., 2000). We used three kinds of merging that correspond to three different taxonomic levels: genus (*Rosa*), section (*Pimpinellifoliae*) and species (*R. persica*). This progressive subdivision scheme will help to see the resolving power of SSRs at different taxonomic levels, from infrageneric to infraspecies relationships. We name these three groups *Ros*, *Pim* and *Rpe*, for genus *Rosa*, section *Pimpinellifoliae* and species *R. persica*, respectively. Accessions from *R.* sect. *Pimpinellifoliae* and *R. persica* were also combined in an artificial unit to study possible gene flows between the two groups. Each STRUCTURE analysis was run three times independently using the Admixture Model and 100,000 Markov chain Monte Carlo (MCMC) repeats with an initial burnin period of 30,000 MCMCs. The number of subpopulations (K clusters) was assessed for ten models involving 1 to 10 subpopulations. We used STRUCTURE Harvester to draw the Evanno plots that served to determinate the most likely number of subpopulations (Evanno et al., 2005; Earl and vonHoldt, 2012). The results from the three independent runs were post-processed to obtain a consensus solution for membership probabilities using Clumpak (Kopelman et al., 2015).

4.4 Results

4.4.1 Numerous alleles to study the genetic diversity of *Rosa* species

The genotyping procedure successfully recovered at least one allele for 77.1% of the 3 040 multiplex PCRs (Appendix C, Supplementary Figure C.1, Additional File 1). Missing data varies greatly from one SSR to another, ranging from 0% (Rog9, RW52D4) to 99% (Rh80). The 32 SSRs used to genotype the 91 (+4 controls) *Rosa* accessions were highly polymorphic with a total number of 1016 observed alleles (Ao), ranging from 1 (Rh80) to 64 (CTG623) alleles per SSR locus (Table 7). The average number of allele (Am) per SSR ranges from 1 (Rh80) to 3.7 (Rw16E19). Effective allele number (Ae) was generally higher than Am and varies from 0.85 (Rw16E19) to 7.3 (RMS082). There were 631 rare alleles, ie represented in less than 5% of the individuals. 34% of alleles were unique to one sample and were therefore not informative. The fraction of new alleles varies from 0% (Rh80, Rw53O21, Rog9) to nearly 60% (Rw16E19, H17C12) in comparison to Liorzou et al. (2016). A total of 178 new alleles were discovered across the 32 SSR loci (Appendix C, Supplementary Table C.2).

4.4.2 Partial taxonomic delineations for recent species relationships

Morphological and biogeographical markers

A set of nine morphological and biogeographic characters was used to assess their ability to discriminate species at the genus level (Additional file 2). Only two clusters were identified (Figure 42A) although the PCA analysis may distinguish more trends based on our knowledge of the genus (Figure 42B). We drew areas that might be linked with the morphological classification. A1 mostly corresponds to subgenera different than subgenus *Rosa*, except for *R. tsingligensis* (TSI01) which is a *Pimpinellifoliae* species. A2 encompasses *Pimpinellifoliae* accessions and overlaps A3 that contains *Rosa* accessions. Most *Caninae* are in A4. A5 includes Asian sections (*Laevigatae*, *Banksianae*, *Bracteatae*) and most of the *Synstylae* species. A1 and A5 approximately correspond to the first k-medoid cluster while A2, A3, and A4 fit the second k-medoid cluster. A clear delineation of subgenera and sections was quite difficult using our set of nine morphological/biogeographical markers. We therefore genotyped all accessions using 32 SSRs to study the extent to which SSRs can distinguish different taxonomic groups in *Rosa*.

Infrageneric relationships

At the genus level, the structure analysis identified three groups (*Ros*1, *Ros*2 and *Ros*3) (Figure 43A; Appendix C, Supplementary Figure C.2A). *Ros*1 encompasses all the *Pimpinellifoliae* accessions while *Ros*2 corresponds mostly to *Synstylae* and *Caninae* accessions. Most accessions from *R.* sect. *Rosa* are shared between *Ros*1 and *Ros*2, displaying substantial percentage of assignment to the two groups (*R. beggeriana* (BEG01, BEG02), *R. amblyotis* (AMB01) are closer to *Ros*1 while *R. majalis* (MAJ03, MAJ04), *R. macrophylla* (MAC03) are closer to *Ros*2). *Ros*3 corresponds to all accessions of *R. persica* that are all well assigned, with membership higher than 90%. The NMDS plot shows the genetic distance between *Rosa* accessions and also spreads the genus in three main parts, although the goodness-of-fit is quite low, as conveyed by a stress value above 0.2 (Figure 43B). Accessions segregate along the Y-axis from top (*Ros*1) to bottom (*Ros*2) with an overlapping region that approximately corresponds to *Caninae* accessions and accessions that were not well assigned to either *Ros*1 or *Ros*2. All accessions from *Ros*3 forms a well delimited group on the left side of the plot. We observed a weak, yet significant, positive correlation between genetic and morphological/biogeographical distances (Mantel test $R^2_M = 0.29 > 0$) (Appendix C, Supplementary Figure C.3).

Table 7: Descriptive statistics and genetic diversity observed in the 91(+4 controls) accessions of *Rosa*. The set of SSR was identical to that of Liorzou et al. (2016), except that we used H17C12 instead of RMS140. No observed, number of observed genotypes; Am, mean number of alleles per individual; Ao, number of observed alleles; Ae, effective number of alleles; rare allele, number of alleles present in <5% of individuals; unique alleles, number of alleles found in only one accession; new alleles, number of alleles that were not observed in Liorzou et al. (2016). Percentages in parenthesis are relative to the total number of alleles for the corresponding SSR.

SSR	LG	No. observed	Missing data (%)	Ao	Am	Ae	Rare alleles	Unique alleles	New alleles
H9B07	1	94	1%	19	1.5	2.3	12 (63%)	10 (53%)	3 (16%)
RMS015	1	91	4%	33	1.6	3.6	26 (79%)	13 (39%)	4 (12%)
RMS070	1	64	33%	37	1.6	5.5	28 (76%)	18 (49%)	16 (43%)
Rw25J16	1	65	32%	58	2.5	4.7	43 (74%)	25 (43%)	6 (10%)
Contig172	2	94	1%	18	2	1.9	9 (50%)	8 (44%)	3 (17%)
CTG329	2	88	7%	12	1.4	1.3	6 (50%)	3 (25%)	1 (8%)
Rh80	2	1	99%	1	1	1	0 (0%)	1 (100%)	0 (0%)
RMS082	2	56	41%	57	2.3	7.3	36 (63%)	24 (42%)	8 (14%)
RMS132	2	93	2%	41	2.6	3.3	22 (54%)	8 (20%)	6 (15%)
RMS147	2	42	56%	25	1.6	3.1	20 (80%)	15 (60%)	4 (16%)
BFACT47	3	91	4%	24	2	2.9	12 (50%)	9 (38%)	3 (13%)
CTG21	3	82	14%	18	1.6	2.2	13 (72%)	6 (33%)	2 (11%)
Rh58	3	78	18%	40	2.5	3.3	18 (45%)	9 (23%)	4 (10%)
RMS144	3	94	1%	24	1.8	4	11 (46%)	5 (21%)	4 (17%)
Rw16E19	3	87	8%	46	3.7	0.8	32 (70%)	17 (37%)	27 (59%)
H20_D08	4	14	85%	11	1.4	3.6	0 (0%)	7 (64%)	4 (36%)
H2F12	4	87	8%	59	3	2.6	43 (73%)	21 (36%)	12 (20%)
Rw53O21	4	93	2%	8	1.6	1.8	2 (25%)	1 (13%)	0 (0%)
Rw55E12	4	94	1%	42	1.9	3.6	30 (71%)	20 (48%)	3 (7%)
H17C12	5	7	93%	5	1.3	2.6	0 (0%)	2 (40%)	3 (60%)
H22F01	5	90	5%	31	1.8	5.3	17 (55%)	6 (19%)	4 (13%)
RMS034	5	88	7%	44	3.6	1.9	24 (55%)	9 (20%)	8 (18%)
RW52D4	5	95	0%	31	3	0.9	20 (65%)	16 (52%)	1 (3%)
CL2980	6	47	51%	29	1.7	5.1	19 (66%)	16 (55%)	7 (24%)
CTG623	6	94	1%	64	2.3	5.3	47 (73%)	26 (41%)	10 (16%)
Rog9	6	95	0%	31	2.1	3.5	18 (58%)	7 (23%)	0 (0%)
Rw22A3	6	88	7%	36	3.3	1.5	15 (42%)	9 (25%)	1 (3%)
H10D03	7	91	4%	47	2.2	3.9	31 (66%)	17 (36%)	9 (19%)
RMS003	7	79	17%	37	1.6	6.7	26 (70%)	18 (49%)	3 (8%)
RMS124	7	51	46%	23	1.3	7	16 (70%)	11 (48%)	2 (9%)
Rw15D15	7	42	56%	27	2.7	2.1	12 (44%)	8 (30%)	12 (44%)
Rw5G14	7	94	1%	38	2.3	3.6	23 (61%)	10 (26%)	8 (21%)
Total		95	-	1016	-	108	631	375	178
Mean		74	22%	32	2.1	3.4	20 (55%)	12 (39%)	6 (18%)

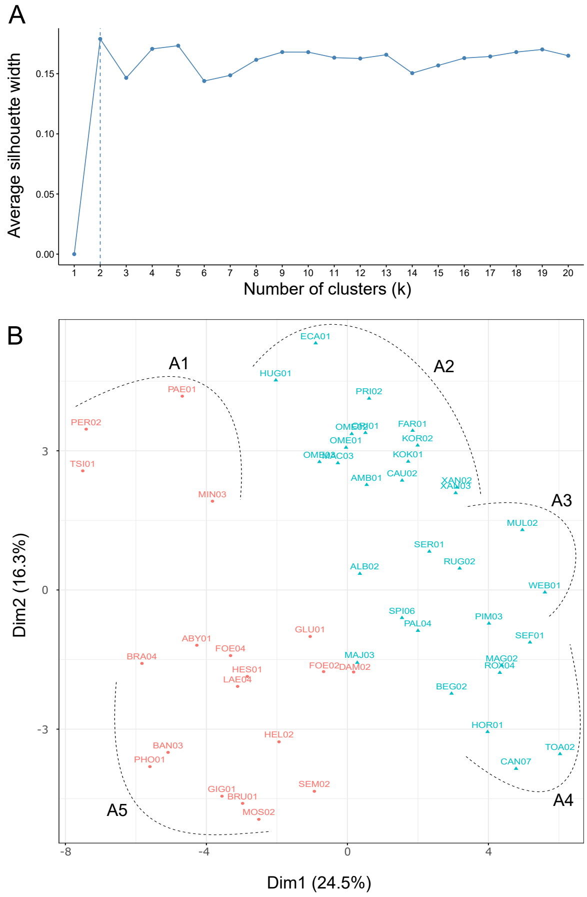


Figure 42: Resolving power of nine morphological and biogeographical characters to distinguish subgroups in the genus *Rosa*. **A**. Silhouette plot indicating the optimal number of clusters to be retained given the morphological/biogeographical distances observed between accessions. **B**. Principal Component Analysis resulting from the partitioning around medoids (PAM) method developed to identify subgroups in the genus *Rosa*. Corail: group 1 and blue: group 2. Areas from A1 to A5 were positioned manually based on knowledge of the genus *Rosa*.

Infrasectional relationships

At the section level, *Pimpinellifoliae* accessions are also structured in three groups (*Pim*1, *Pim*2, and *Pim*3) (Figure 44A; Appendix C, Supplementary Figure C.2C). Pim1 approximatively corresponds to *R. spinosissima*-like accessions and North Asian *Pimpinellifoliae* species. *Pim*2 is the smallest group with most of its accessions sharing large membership to either *Pim*1 or *Pim*3. *Pim*2 only include bright-yellow flowered species. *Pim*3 gathers yellow flowered species as well as Asian *Pimpinellifoliae* from the *R. sericea* complex. The genetic distance analysis also segregates the *Pimpinellifoliae* section in subgroups (Figure 44B). The X-axis separates *Pimpinellifoliae*-like species on the left (typical yellow-flowered species) from section *Rosa*-like species on the right. Among other species, the *Rosa*-like group encompasses (1) *R. spinossissima* (syn *R. pimpinellifolia*), the type species of the *Pimpinellifoliae*, that was shown as an intersectional hybrid involving

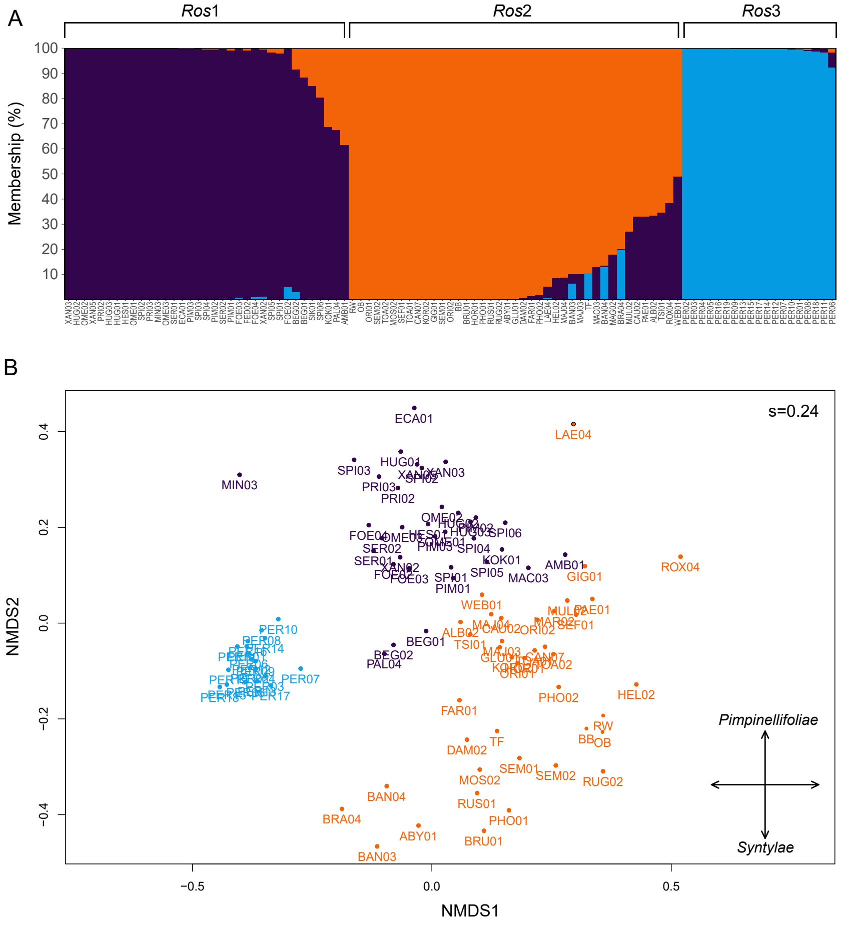


Figure 43: Genetic distance and structure at the genus level. **A**. Proportional membership of each accessions in the genetic clusters inferred by STRUCTURE with K = 3. Each individual is represented by a vertical bar, and the length of each bar indicates the membership probability in each cluster. **B**. Non-metric multidimensional scaling plot showing the Dice's genetic distance between samples at the genus level (genus *Rosa*). s refers to the stress value ($0 < s < 1$). Accession labels and points are colored according to their main cluster as found in the structure analysis.

R. sect. *Rosa* and remaining members of *R.* sect. *Pimpinellifoliae*, and (2) *R. farreri*, another *Pimpinellifoliae* accession that is supposed to belong to *R.* sect. *Rosa*. The Y-axis divides the *Pimpinellifoliae* section geographically, with Middle Eastern accessions at the top (all *R. foetida* accessions, *R. hemisphaerica* var. *rapinii*, *R. kokanica*, *R. spinosissima* var. *altaica*) and East Asian accessions at the bottom (accessions from the *R. sericea* complex).

Gene-flows between section *Pimpinellifoliae* and subgenus *Hulthemia*

There was no gene flow between section *Pimpinellifoliae* and subgenus *Hulthemia* as shown in Figure 45. The artificial group made by aggregating accessions from section *Pimpinellifoliae* and subgenus *Hulthemia* (*R. persica*) was structured in two well delimited groups (g1 and g2) that encompass all the *Pimpinellifoliae* accessions (g1) and all the *R. persica* accessions (g2) (Figure

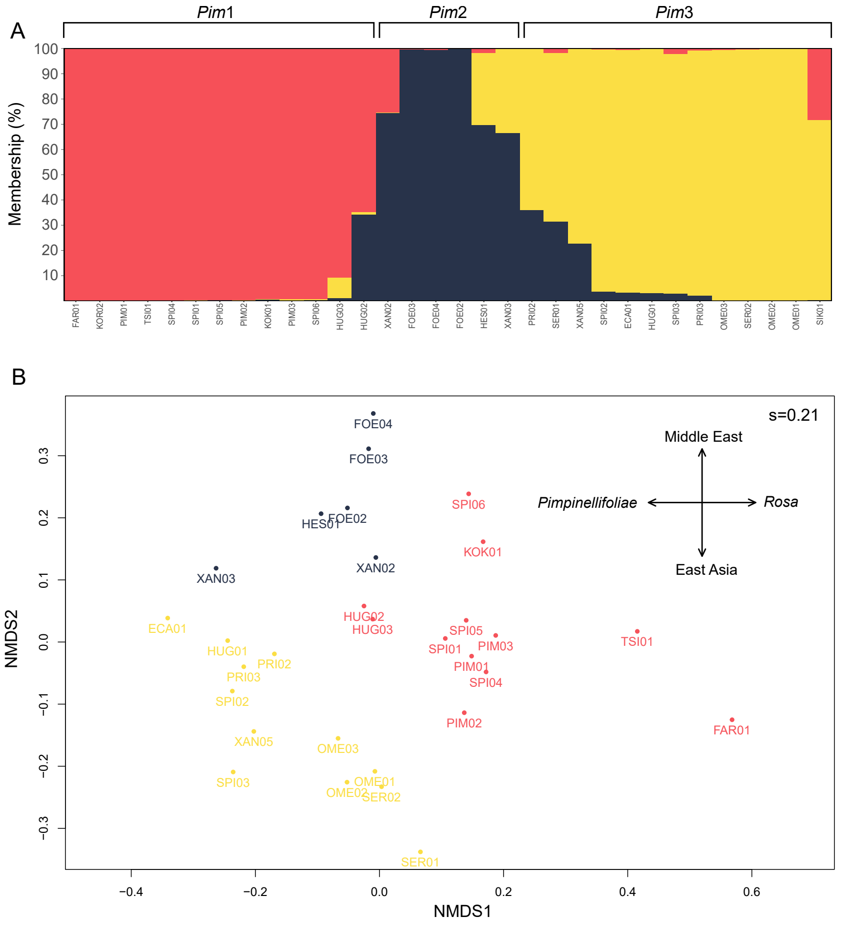


Figure 44: Genetic distance and structure at the section level. **A**. Proportional membership of each accessions in the genetic clusters inferred by STRUCTURE with K = 3. Each individual is represented by a vertical bar, and the length of each bar indicates the membership probability in each cluster. **B**. Non-metric multidimensional scaling plot showing the Dice's genetic distance between samples at the section level (*Pimpinellifoliae*). s refers to the stress value ($0 < s < 1$). Accession labels and points are colored according to their main cluster as found in the structure analysis.

45A; Appendix C, Supplementary Figure C.2B). All accessions have a membership above 90% to either g1 or g2. The genetic distance analysis also divided the genus in two distinct groups with no overlap between section *Pimpinellifoliae* and subgenus *Hulthemia* (Figure 45B).

4.4.3 Conflicting relationships between geography and genetics in *Rosa persica*

The resolving power of SSRs was assessed at the species level, which corresponds to the basic unit of the taxonomic rank. We studied the genetic structure of 19 accessions of *Rosa persica* from diverse locations in the Middle East region. Most of the samples were collected along with precise spatial coordinates. We first observed that the spatial distance does not correlate with the genetic distance, hinting at possible flows between accessions (Appendix C, Supplementary Figure

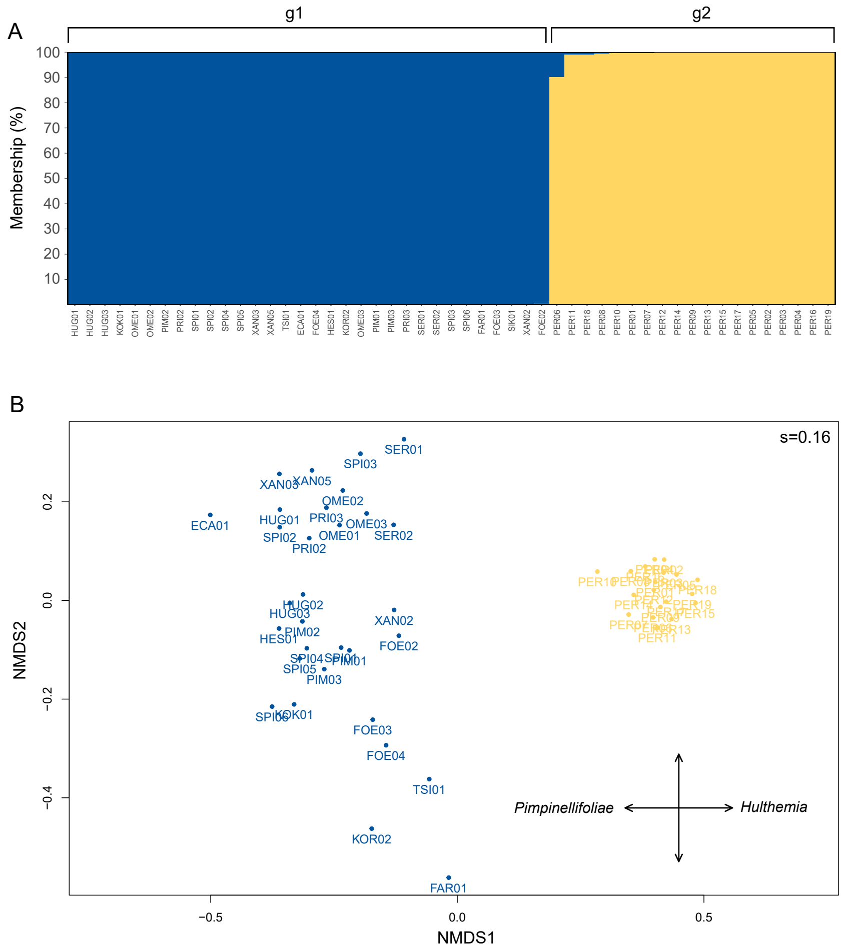


Figure 45: Genetic distance and structure inside the artificial operating taxonomic unit *Pimpinellifoliae-Hulthemia*. **A**. Proportional membership of each accessions in the genetic clusters inferred by STRUCTURE with K = 3. Each individual is represented by a vertical bar, and the length of each bar indicates the membership probability in each cluster. **B**. Non-metric multidimensional scaling plot showing the Dice's genetic distance between samples at the section level (*Pimpinellifoliae*). s refers to the stress value ($0 < s < 1$). Accession labels and points are colored according to their main cluster as found in the structure analysis.

C.4). We further investigated the genetic structure of the set and identified seven groups (from *Rpe*1 to *Rpe*7) (Figure 46A; Appendix C, Supplementary Figure C.2D). Most of the group contain accessions that are well assigned (membership > 80%) except *Rpe*2 and *Rpe*3 that show substantial fractions of two subgroups. *Rpe*1 encompasses four accessions from a same population sampled nearby Mashhad, Iran. The genetic distance analysis failed to identify a consistent explanation to the two axis, although some accessions that share a similar genetic structure are found close to each other on the NMDS plot (see accessions from *Rpe*1, *Rpe*2, *Rpe*3 and *Rpe*4) (Figure 46B). Reliefs and specific landforms can create natural fences that may isolate populations. Therefore, we constructed a hill-shaded map (Figure 47) to (1) investigate why species that are spatially close to each other are ultimately found genetically distant and (2) see if the genetic structure of *R. persica* accessions can be explained by specific landforms. We found no clear evidence for explaining the genetic structure observed in *R. persica* accessions, although some isolated accessions seem to have their own genetic structure (PER16, PER17, PER19).

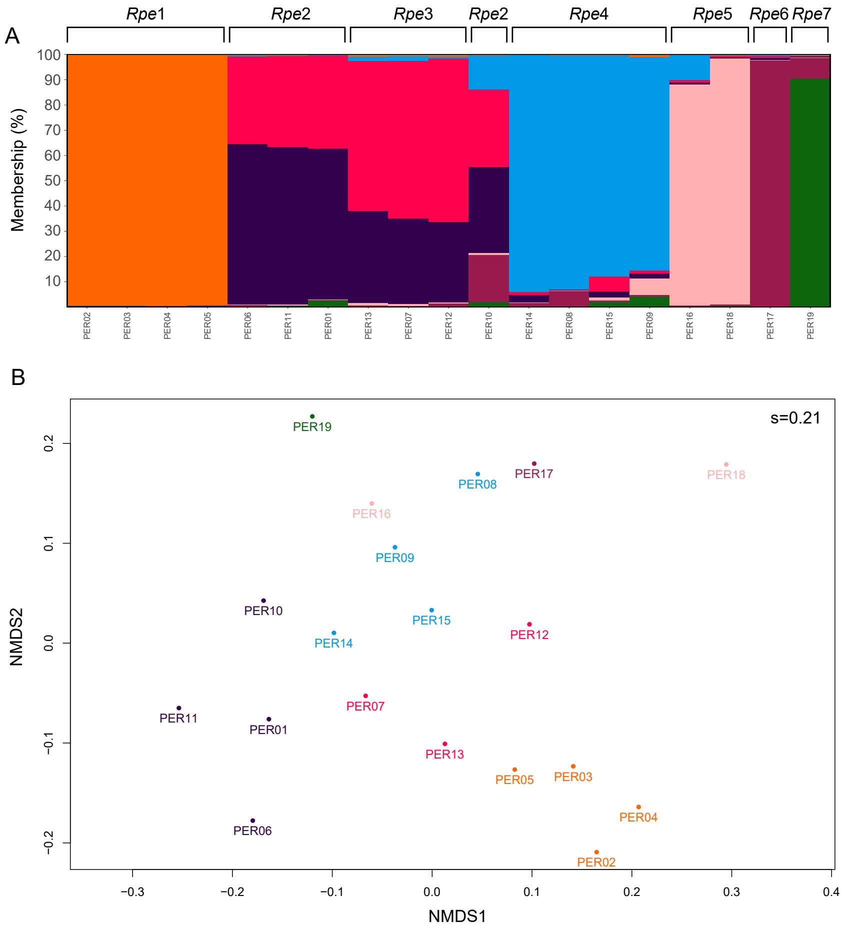


Figure 46: Genetic distance and structure at the species level. **A**. Proportional membership of each accessions in the genetic clusters inferred by STRUCTURE with K = 7. Each individual is represented by a vertical bar, and the length of each bar indicates the membership probability in each cluster. **B**. Non-metric multidimensional scaling plot showing the Dice's genetic distance between samples at the species level (*Rosa persica*). s refers to the stress value ($0 < s < 1$). Accession labels and points are colored according to their main cluster as found in the structure analysis.

4.5 Discussion

4.5.1 The use of SSRs to identify consistent groups at different taxonomic ranks

The genotyping successfully identified a large number of alleles (1016) over only 32 loci, a yield that would have probably not been possible with any other marker. Although one third of the alleles were specific to only one accession, we identified 178 new alleles that complement the study

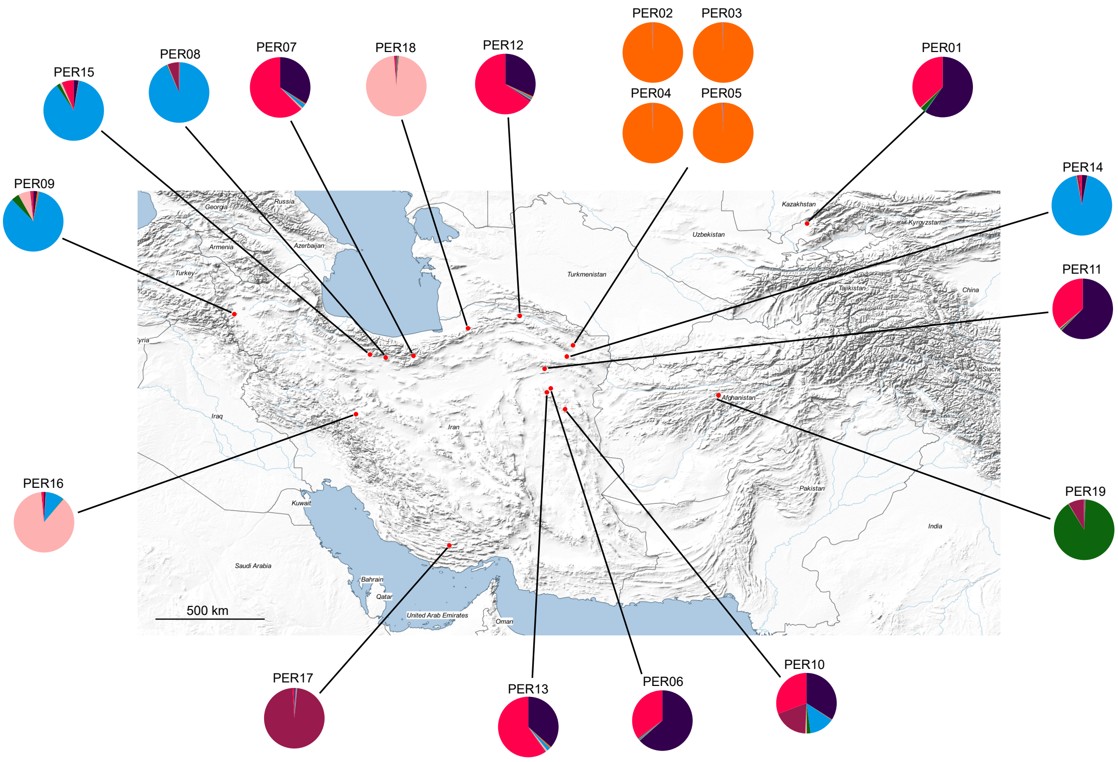


Figure 47: Geographic distribution of 19 samples of *Rosa persica* in the Middle East. The hill-shaded map was obtained in QGIS based on datasets from Natural Earth. The proportional membership of each accession as obtained with STRUCTURE is also plot for a better visualization of the relation between geographic distance and structure analysis.

previously made on cultivated material (Liorzou et al., 2016). We produced genotyping data that will help to bridge the gap between the study of cultivated accessions and their wild counterparts since we genotyped our wild accessions with the same set of SSR markers as Liorzou et al. (2016). The genotyping information that we developed in this study will pave the way for further in-depth analyses that would consider both wild and cultivated rose material, as it has been done in other groups (Zhang et al., 2017; Maraci et al., 2018). Such approaches would enable to study the genetic contribution of wild species to the development of ornamental cultivars, and roses are an appropriate model to do so.

In our attempts to delineate more precisely the boundaries of sections and species in the genus *Rosa*, we only observed quite vague groups, especially at the infrageneric level (Figure 42B and 43). Both morphological/biogeographical and genetic distance analyses resulted in mitigated segregations. There are several explanations for such poor resolution at the infrageneric level. First, our sampling was biased towards section *Pimpinellifoliae* and subgenus *Hulthemia* which might have unbalanced the analysis and the resulting two-dimensional distance plots. Second, the assignment of morphological/biogeographical characters to our samples was based on literature and may not reflect the intraspecific variability that we sometimes detected in genetic analyses. Therefore, we observed a weak correlation between the genetic and the morphological/biogeographical distances. In addition, our set of just nine morphological/biogeographical characters might be insufficient to summarize the diversity in *Rosa*. Therefore, SSR markers may not represent an ideal choice for studying species relationships at the infrageneric level. In addition, SSR loci are generally supposed to evolved fast (Kruglyak et al., 2000; Bhargava and Fuentes, 2010), which may by chance lead to identical allele in different species, although this proximity does not reflect phylogenetic relationships but likely corresponds to homoplasic relationships. Traditionally, SSR allele information is extracted only through recording fragment length. However, indels may occur in SSR flanking regions leading to allele length identities while allele structures are different. This phenomenon is called size homoplasy and has been reported prominently in angiosperms (Šarhanová et al., 2018). Therefore, non-distance-based methods may still perform best to recover a comprehensive evolutionary history at the genus level.

At shallower taxonomic rank, the use of SSR may enable to identify trends with a biological meaning. At the section level, our analysis clearly demonstrated a two-scale segregation. First, *Pimpinellifoliae* accessions segregated between two sets: the *Pimpinellifoliae*-like species and the *Rosa*-like species (Figure 44). We clearly showed that *R. farreri* is outside of the core *Pimpinellifoliae* section and may better belong to section *Rosa*, as suggested in (Roberts, 1977; Fougère-Danezan et al., 2015). We also observed a cluster of *R. spinosissima* (syn. *R. pimpinellifolia*) accessions in between the two sets which is in line with previous analyses that identified this species as an intersectional hybrid between *R.* sect. *Pimpinellifoliae* and *R.* sect. *Rosa* lineages (Chapter 3). The position of *R. foetida* was also quite intermediate, as observed in other study (Bruneau et al., 2007; Fougère-Danezan et al., 2015). The polyphyly of section *Pimpinellifoliae* may be explained by its proximity with section *Rosa*, with possible intersectional hybrids. Second, *Pimpinellifoliae* accessions segregated geographically between Middle Eastern species and East Asian species, with yellow-flowered species closer to Anatolia and the Middle East, and species from the *R. sericea* complex closer to the Himalayas and western China mountains in East Asia.

At the species level, the use of SSRs enabled to identify different alleles in closely related accessions, as well as to find different genetic structures (Figure 46). However, it was difficult to link the observed differences with biogeographic data (Figure 47). A denser sampling with many more individuals per group might have led to more substantial results.

4.5.2 On the relationships between section *Pimpinellifoliae* and subgenus *Hulthemia*

There were several arguments for considering that section *Pimpinellifoliae* and subgenus *Hulthemia* could have shared similar genetic features. Both groups include species with bright yellow flowers, a color that is found nowhere else in the genus *Rosa*. The distribution of their respective species overlap significantly in some areas, especially in the Middle East. Recent phylogenetic analyses identified close phylogenetic relationships between lineages leading to section *Pimpinellifoliae* and subgenus *Hulthemia* (Chapter 3). Despite of all these similarities, using SSR markers we demonstrated that there was no gene flow between section *Pimpinellifoliae* and subgenus *Hulthemia*. The color similarity is likely to be attributed to their environment that may be somewhat arid, with few ecological niches for pollinating insects. By developing flowers with attractive color, yellow-flowered species may increase their chance to be pollinated. This is especially true for *R. persica* that thrive in arid, sandy and stony places and developed a macula, ie a dark blotch in the middle of the flower, which is an effective strategy for appealing to bees and butterflies that are attracted by contrasted colors (Lebel et al., 2017; Hirota et al., 2018).

The phylogenetic proximity between section *Pimpinellifoliae* and subgenus *Hulthemia* may be impossible to study using genotyping method. Both of their lineages branch deeply in the most recent nuclear phylogenetic analyses which limit the possible detection of gene flows using rapid evolving SSRs. In a plastid-based divergence time analysis, Fougère-Danezan et al. (2015) identified an ancient phylogenetic relationship between lineages that led to *R. persica* and *R. minutifolia*, which contrasts with the results obtained in the recent phylogenomic study (Chapter 3) and that show a phylogenetic proximity between the main *Pimpinellifoliae* lineage and that of the subgenus *Hulthemia*. Moreover, aside from being arid-adapted, *R. persica* and *R. minutifolia* do not share a similar distribution. Fougère-Danezan et al. (2015) developed a biogeographical explanation for the endemicity of *R. minutifolia* in the southern regions of North America. *R. minutifolia* would correspond to the remain of a first colonization and expansion of the genus prior to the *Grande Coupure*. Putting together the results from phylogenomics and divergence analysis, we could hypothesize that *R. persica* also derived from a first colonization and expansion of the genus. The divergence between subgenus *Hulthemia* and the main *Pimpinellifoliae* clade may have occurred before the *Grande Coupure* (around 34 MYa). *R. persica* ancestors might have survived the rapid decrease of temperature due to their southernmost distribution. Ancestors of the main *Pimpinellifoliae* lineage may have been drastically selected after the *Grande Coupure* and led to the development of the many *Pimpinellifoliae* species that we now today. Therefore, *R. persica* cannot be considered as a living fossil of the genus *Rosa*, or as an ancestor common to all wild roses. *R. persica* rather derived from an ancient lineage already included in the genus *Rosa* and that was shared with the *Pimpinellifoliae* species. This early lineage led to some species with arid-adapted features.

4.6 Conclusion

Phylogenetic relationships between very closely related species are difficult to assess which hamper the recovery of a comprehensive evolutionary history of the group. Using genotyping markers instead of ubiquitous DNA sequences may bring more variations to (1) study sister taxa at the boundaries between species and populations, (2) study the genetic contribution of the wild pool to cultivated materials. We demonstrated the resolving power of SSRs for studying intrasectional and intraspecies relationships. In the *Pimpinellifoliae* section, two segregating patterns were observed and could have been explained with both biogeographical and phylogenetic arguments. At the intraspecific level, SSRs enabled to structure the subgenus *Hulthemia* but there was no clear link with biogeography. We advocate to use many more individuals to improve the resolution and the value of such studies. We did not observe any gene flow between section *Pimpinellifoliae* and subgenus *Hulthemia* which may be explained by a too deep divergence in the evolutionary history of the genus *Rosa*. This demonstrates the value of combining phylogenetic and population-like genetic studies for describing the evolutionary history of *Rosa*.

4.7 Additional file and availability of data and material

Additional File 1. Allele table as output by GeneMapper v4.1 after the filtering steps (Excel file).

Additional File 2. Morphological and biogeographical characteristics of the 91 accessions based on literature (Excel file).

The two aforementioned Additional Files and sample material are available upon request to the GDO team.

4.8 Author's contribution

KD performed the analyses and drafted the manuscript. EM extracted SSR allele data in GeneMapper and identified new alleles. EM collected morphological and biogeographical data in literature. ZK collected samples of *R. persica* in Iran and help with the DNA extractions. AC designed the genotyping experiment and provided valuable advice for using GeneMapper. FF and VM proofread the manuscript.

4.9 Acknowledgments

The authors thank A. Bruneau (University of Montréal) for providing samples of *Rosa* used in her past studies. We also thank the staff of Arboretum du vallon de l'Aubonne, The Arnold Arboretum of Harvard University, Botanical Garden of the University of Bonn, Cambridge University Botanic Garden, Conservatoire Botanique National de Franche-Comté, Botanical Garden of the University of Strasbourg, Botanical Garden of the City of Nantes, The Botanical Garden of the Finnish Museum of Natural History, Ferdowsi University of Mashhad, Botanic Garden Meise, The Morton Arboretum, Quarryhill Botanical Garden, Royal Botanic Garden of Edinburgh, Roseraie du Val-de-Marne, Roses Loubert, Rancho Santa Ana Botanic Garden, Unité Expérimentale de la Villa Thuret, Wageningen University & Research, for preparing and/or providing *Rosa* samples. The authors are grateful to the Gentyane genotyping platform, Clermont-Ferrand, France for assistance with SSR genotyping.

5 General discussion and perspectives

This thesis focused on overcoming the difficulties associated with the phylogenetic study of complex taxonomic groups through the case study of the genus *Rosa*. One of the main challenge was to develop a comprehensive framework for the phylogenetic study of this genus, given its inherent complexity and the limitations of current phylogenetic methods. In *Rosa*, the large number of species associated with their ease to spontaneously hybridize at odd and even ploidy levels, sometimes using asymmetrical meiosis, make this genus a challenge for phylogeneticists. Indeed, the overall aim of any phylogenetic analysis is to provide a robust and comprehensive evolutionary framework for the considered group. This implies that the resulting phylogeny must be (1) representative of the taxonomic group, (2) robust, and (3) able to appraise the processes of evolution that shaped the group. However, there is a substantial gap between the three aforementioned expectations and the current phylogenetic methods when large and complex taxonomic groups are considered. Most phylogenetic methods were indeed developed for simple case scenarios involving matrices with few genes and few taxa. Achieving a robust phylogeny on a large number of taxa often implies using a large number of molecular sequences through phylogenomics. This is especially true in taxonomic groups involving numerous closely related species, such as *Rosa*. Accumulating a large number of molecular sequences is the hope of obtaining enough informative variations to recover a robust phylogeny. However, dealing with large arrays of sequences implies the use of substantial computational resources to apply realistic DNA substitution models. If one wants to add more sophisticated parameters to the evolutionary models, such as the consideration of reticulate patterns due to hybridization, then the computational burden of phylogenomics simply becomes intractable. Taking the genus *Rosa* as an example, we provided a general framework to address the reticulate phylogenetic relationships within a large group of intertwined, closely related species (Figure 48). Our stepwise approach enabled to significantly improve the phylogeny of *Rosa* by (1) obtaining a robust backbone phylogenetic hypothesis and (2) considering both hybridization and polyploidy in the evolutionary history of the genus. Here, we discuss the implication of such approach for phylogenetics in large and complex taxonomic groups, as well as for the study of the evolutionary history of *Rosa* and its use in breeding.

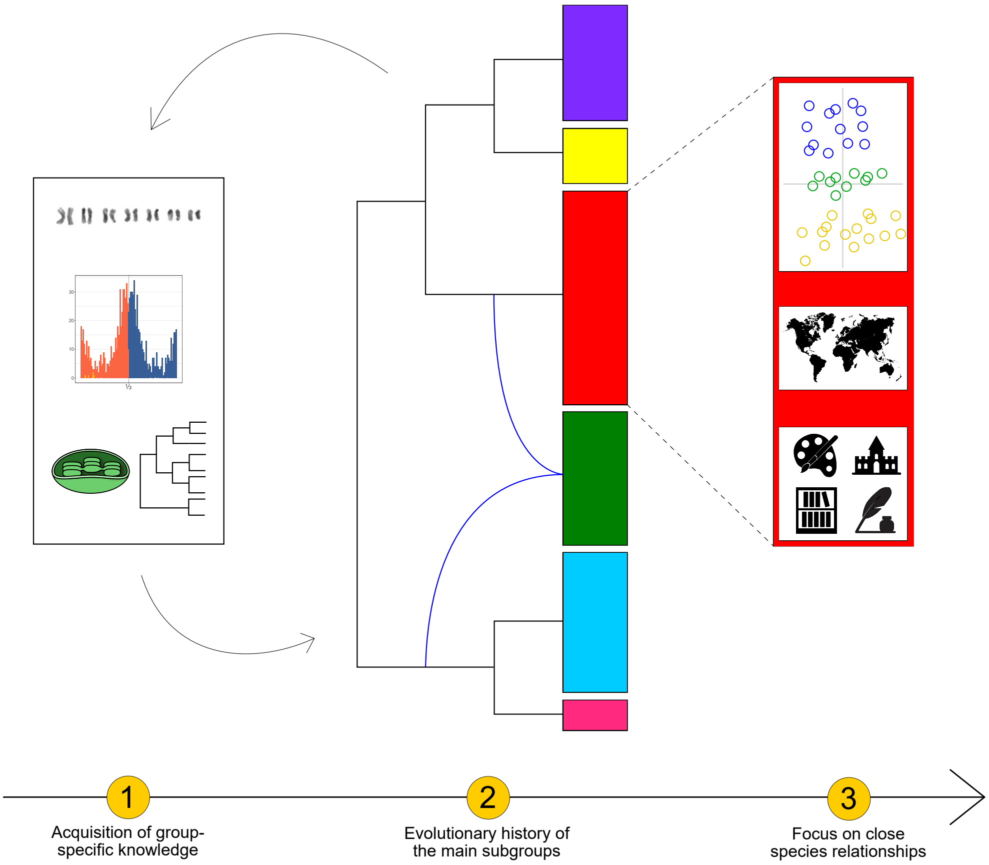

1
Acquisition of group-specific knowledge
2
Evolutionary history of the main subgroups
3
Focus on close species relationships

Figure 48: Stepwise procedure to study the species relationships within large and complex taxonomic groups. The procedure includes three main steps. **Step 1** Acquisition of knowledge specifically related to the taxonomic group. This encompasses karyotyping, ploidy estimations, plastid phylogenies, phenotypic data, etc. . . . This steps enables to better know the group that will be studied by identifying similar individuals, pinpointing odd specimens and providing a global and simple bifurcating pattern in the case of plastid phylogeny. **Step 2** Study of the phylogenetic relationships between the main subgroups/sections. This implies to recover a robust backbone phylogeny of diploids further brighten up with reticulations between subgroups. **Step 3** Focus on close species relationships. As we get more interested in closely related species, the amount of genetic similarities/exchanges is expected to be high causing the relationships between taxa to be blurred and very reticulated. Traditional phylogenetics can no longer bring information at this scale. It is therefore necessary to use other methods such as the study of phenetic relationships, biogeography and ecology, and historical data in the case of plants that were selected by people for centuries.

5.1 A global framework to address phylogenetic relationships within large and complex taxonomic groups

5.1.1 Wise selection of phylogenetic markers

In Chapter 2, 1856 single-copy orthologous tags ($\mathrm{SCO_{Tag}}$s) were developed to study the phylogenetic relationships of *Rosa* through phylogenomics. $\mathrm{SCO_{Tag}}$s combine numerous advantages for phylogenetic study and overcome most of the shortcomings associated with the use of universal phylogenetic markers. By targeting informative single-copy loci with unique reciprocal relationships, we favored the selection of orthologous sequences capable of addressing species relationships at the shallowest taxonomic ranks (see Figure 25). One of the main advantages of such approach is that the set of phylogenetic markers is fully dedicated to the taxonomic group that is considered. In addition, the phylogenetic utility of $\mathrm{SCO_{Tag}}$s was appraised through a thorough analysis of phylogenetic informativeness and conflicting signals (Figure 27 and Figure 28), while most studies aiming at developing phylogenetic markers do not characterize their datasets to such an extent (Cabrera et al., 2009; Duarte et al., 2010; Liston, 2014; Chery et al., 2017; Choi et al., 2019) and can even lead to poor supports despite using multiple nuclear loci (Yang and Davis, 2017). A similar approach to ours was developed by Granados Mendoza et al. (2015) on Hydrangea, although it recovered far less sequences for phylogenetic studies. The method that we developed for $\mathrm{SCO_{Tag}}$s mining can be applied to any taxonomic group that have an available reference genome along with whole genome sequencing read sets for some member of the group. We stress that a careful selection of loci maximizes the chances to recover informative sequences from most group members and therefore to obtain a robust phylogeny. The method that we provided for $\mathrm{SCO_{Tag}}$ identification strongly focuses on selecting genome-scale informative variations. This contrasts with other approaches that rely on the identification of phylogenetic sequences solely in transcriptomic, exonic datasets, and ultra-conserved elements (UCEs) (Faircloth et al., 2012; Mandel et al., 2014; Wickett et al., 2014). Such designs may be suitable to studying phylogenetic relationships at medium to deep taxonomic ranks (Bragg et al., 2016). However, dealing with very close species relationships require to target variations that might only exist in non-coding parts of the DNA sequences (Sang, 2002; Li et al., 2017). In addition, we argue that targeting complete sequences rather than collecting single nucleotide variants (SNVs) or Restriction-site Associated DNA sequencing (RADseq) data, provides unbiased material to accurately model evolutionary relationships. Indeed, phylogenomics based on variant calling may be subjected to ascertainment bias which in turns can distort branch lengths estimations (Lewis, 2001; Leaché et al., 2015). Ascertainment bias therefore needs to be corrected, but all types of correction have significant limitations (Tamuri and Goldman, 2017). Our $\mathrm{SCO_{Tag}}$ identification procedure ensures the targeting of sequences with the right balance between conserved regions, to facilitate the recovery of sequences from all group members, and more variable regions to distinguish groups of individuals close to each other. However, we acknowledge that $\mathrm{SCO_{Tag}}$s face some limitations. First, although we tried to limit the selection of duplicated sequences though the use of single-copy genes and reciprocal best BLAST searches, we could not fully guarantee the absence of hidden paralogy among our set of SCO$_{\text{Tag}}$s (Figure 22). Hidden paralogy between a set of SCO$_{\text{Tag}}$ sequences might be detected through a careful study of non-synonymous/synonymous rates of substitutions (Yu et al., 2017). However, such bias may be already visible for SCO$_{\text{Tag}}$ whose phylogenetic informativeness is out of range on the PI profiles. Another drawback of SCO$_{\text{Tag}}$ is that it is impossible to phase their alleles. Allele A in SCO$_{\text{Tag}}$ 1 might not originate from the same sub-genome as Allele A in SCO$_{\text{Tag}}$ 2. This may be problematic to study genomic exchanges at the chromosome level.

Although tens of SCO$_{\text{Tag}}$s were highly informative, we observed that most of the remainders correspond to highly conserved loci. Such levels of conservation might not be useful to unveil close species relationships using DNA evolutionary models. Therefore, only a fraction of the 1856 SCO$_{\text{Tag}}$s can be used for phylogenetics. However, developing such conserved tags has several other advantages. SCO$_{\text{Tag}}$s can indeed be seen as milestones in genomes and can therefore serve to compare the synteny between WGS assemblies (Appendix A, Supplementary Figure A.1). Evolutionary biologists could therefore benefit from the conserveness of SCO$_{\text{Tag}}$s to appraise genomic gains and losts between different genomes inside the taxonomic group. In addition, we demonstrated that a substantial part of SCO$_{\text{Tag}}$s were still informative enough to decipher the origin of closely related taxa in the Chinenses section. We demonstrated that SCO$_{\text{Tag}}$s represent an excellent resource for chromosome painting (Figure 31). In addition, we think that SCO$_{\text{Tag}}$s can also serve to anchor scaffolds from WGS assemblies of any related specimen onto the reference genome of the group.

5.1.2 Phylogenetics at the time of Next-Generation Sequencing techniques

Investigating the evolutionary history of a large and complex taxonomic group implies to use a sufficient amount of data to recover supported phylogenetic relationships. This results in large arrays of taxa/sequences. Generating such datasets is out of reach of classic Sanger sequencing and necessitate the use of high throughput sequencing techniques. The selection of the most appropriate NGS technique typically depends on many factors such as the size of region of interest, the robustness with challenging samples, and the price. For large taxonomic group that are distributed throughout vast geographical areas, the access to fresh quality samples is often impossible. Most samples must be dried before sending and the cheapest procedure consist in using silica gel. Although this enable to preserve DNA in a reasonable way, the yield and quality of DNA extracts are lowered. Depending on the species, tissue material and storage conservation, we observed that DNA solution may contain a lot of small DNA fragments (<1kb). The selected NGS technique must therefore accommodate such conditions. In our experiments, we used microfluidic PCRs followed by multiplex Illumina sequencing to obtain million reads for each sample. Amplicon sequencing has many advantages for shallow-scale phylogenetics, including (1) recovery of intermediate-size loci (between 200bp and 600bp) (2) accommodation of degraded samples (3) short data assembly time and (4) cost-effectiveness. However, we acknowledge that amplicon sequencing requires a cumbersome development time to select the best possible target loci which in turn requires to access genome-wide data (WGS reads). In addition, PCR bias and artefacts resulting in chimeric sequences may alter the quality of the amplicon sequencing (Krehenwinkel et al., 2017). In our case, thorough filtering steps were performed during amplicon assemblies to remove chimeras. It is also worth mentioning that our $\mathrm{SCO}_{\mathrm{Tag}}$s are quite conserved, at least in their extremities, and do not present long repetitive structures that would hamper correct amplification and sequencing. An alternative to amplicon sequencing would be hybrid enrichment (Lemmon and Lemmon, 2013). However, this technique generally requires to target conserved regions; so hundreds of loci might be necessary to gather sufficient information for unveiling infrageneric species relationships. Generating such large datasets involves to be able to analyze the resulting tremendous amount of data. Such analysis may be doable for simple case scenario of diploid taxa without gene flows, but becomes intractable for more complex taxonomic groups such as *Rosa*. In addition, hybrid enrichment requires a substantial DNA starting quantity (~500 ng) (Miyazato et al., 2016; Nikolov et al., 2019) while amplicon assays are able to work with much smaller quantities of input DNA, often down to 10 ng (McKain et al., 2018). This is a key advantage of amplicon sequencing for projects that rely on degraded plant material (herbarium, silica dried material conserved for long times) as it was the case in our study. Since DNA is supposed to be largely degraded, high weight DNA molecules may be more difficult to extract. Targeting small DNA fragments (200-600 bp) can be achieved with PCR-based methods in a more cost-effective manner than with hybrid enrichment that is better suited to recovering longer DNA fragments.

5.1.3 Reducing the complexity of the problem

The framework that we developed for *Rosa* phylogenomics is geared toward reducing the problem's complexity. First, we circumvent costly and tedious whole genome sequencing and assemblies by selecting well-distributed tags over the seven pseudomolecules of the *Rosa* genome. Then, we did not fall into the trap of including all possible genomic data in our analyses considering that the true overall phylogenetic signal would arise from the mass and conceal the noise of each tag. Instead, we spend much effort to carefully select our sequences both to gather the most informative regions in the least number of sequences, thus improving the cost-effectiveness of our study. We therefore demonstrated that although we represented the chloroplastic genome with only four tags (total alignment length: ~2000 bp), we still obtained a well resolved phylogeny (Figure 34). Compared to similar phylogenetic studies in terms of gene/taxon sampling with less robustness (Fougère-Danezan et al., 2015; Wang et al., 2016), we have taken care to (1) select truly informative and not only variable areas, (2) thoroughly inspect the alignments for possible misaligned areas and correct them, (3) search and apply the best model of evolution to each sequence, (4) consider the informative indels.

Reducing the complexity of the problem was also achieved during the phylogenetic analyses. We both estimated the ploidy level of each accession based on SNP frequencies in sequencing data (Figure 33), and we pinpointed putative hybrid or mislabeled specimens using a plastid phylogeny

(Figure 34). This approach was of great value since it enabled to acquire more knowledge on our taxa sampling. We were then able to distinct simple taxa (non-hybrid diploids) from more complex taxa (putative hybrid diploids, (allo)polyploids). The greatest challenge was to combine our large taxon sampling with allelic information since we refused to conceal allelic variations into consensus sequences. The development of a backbone phylogeny using only putative non-hybrid diploid specimens provided unprecedented insights into the organization of the genus *Rosa* using tens of nuclear sequences (Figure 35). Taxa from this backbone tree further served as placeholders for the phylogenetic analysis of the set of complex taxa. This enabled to identify groups of polyploids in split networks as well as cross comparisons with plastid phylogenies. However, we went beyond the simple analysis of discrepancies between nuclear and plastid phylogenies by developing hybrid networks (Figure 38). These hybrid networks brought more information than split networks by (1) identifying parental lineages and (2) providing the genome fraction that is inherited from each parental lineage. Again, we had to reduce the problem complexity by selecting putative parental lineages and pointing the possible hybrid specimens out. Proceeding this way, we were able to consider both hybridization and ILS in the computation of hybrid networks. All the biases that we had to take for constructing hybrid networks were based on the results obtained in the first analyses (ploidy estimation, plastid phylogenies), thus limiting the introduction of preconceived facts from literature. Many phylogenetic studies on large taxonomic groups indeed rest on literature to gather information on ploidy level or hybrid origin of certain species (Fougère-Danezan et al., 2015; Kamneva et al., 2017). In multiploid taxonomic groups, where misidentification can be even more frequent, our stepwise approach without a priori can be of great value. Finally, although some step wise approaches were already developed in other taxa, they usually resort to either few species (Kamneva et al., 2017; Díaz-Pérez et al., 2018) or few gene sequences (Marcussen et al., 2012, 2015; Cai et al., 2012; Brassac and Blattner, 2015). To the best of our knowledge, we developed the first phylogenetic studies that provides solutions for (1) obtaining a robust backbone phylogeny and (2) study sample-specific hybridization and polyploidy, using large arrays of taxa and sequences at shallow phylogenetic scales.

5.2 Beyond trees and networks

As far as we know, there does not exist any phylogenetic method that would be able to provide an all-in-one solution for inferring phylogenetic relationships in large and complex groups of closely related species. Therefore, the approach must be stepwise, include different kind of analyses (ploidy estimation, plastid and nuclear phylogenies, small-scale hybridization networks) and be completed by methods that go further than the basic phylogenetic frame (Figure 48). Phylogenetics can provide general patterns (trees, networks) to identify the main subgroups within a taxonomic group, as well as their evolutionary relationships. However, at the boundaries between species and populations, phylogenetics shows its limits and do not provide sufficient resolution to unveil very close species relationships. Therefore, a better characterization of the relationships between closely related taxa must involve the use of additional strategies, borrowed from population genetics. In Chapter 4, we assessed the use of SSRs at different taxonomic levels of the genus *Rosa*. We observed that intrasectional relationships could be appraised through the use of SSR (Figure 44 and Figure 46). However, this would require a much larger taxon sampling to hopefully observe gene flows between groups of individuals. Expanding concepts from population genetics to higher taxonomic levels (section, genus) already proved to be useful to address the species-populations continuum in several genera (Zanella et al., 2016; Wang et al., 2017b; Zhang et al., 2018b). Another way to access genetic diversity at a broader scale than in Chapter 4 consists in using Restriction-site Associated DNA sequencing (RADseq) (Miller et al., 2007; Baird et al., 2008) which has several advantages over SSRs (Lemopoulos et al., 2019) or SNP arrays, especially since there is a reference genome for *Rosa* to map RAD loci. First, RADseq provides an excellent genome-scale resolution compared to the 32 SSRs used in chapter 4. RADseq generally yield about 50K SNP positions, which is however less than the densest SNP array (600K) (Minias et al., 2019). Second, the automation of a bioinformatics pipeline would circumvent the manual and cumbersome reading of peaks on the SSR electropherograms. Third, RADseq is a very cost-effective Genotyping-By-Sequencing (GBS) method since it is two times less expensive than SNP arrays ($35/sample versus $59/sample (You et al., 2018)), but still more expensive than genotyping 32 SSRs per individual ($12/sample (personal estimation)). Accessing read depth in NGS datasets can also provide insights into allele frequencies and thus ploidy levels. Several studies also reported the value of RADseq to study relationships between polyploid taxa (Brandrud et al., 2017; Kinosian et al., 2019). Although RADseq studies involve more extensive processing compared to SNP arrays, they are also more flexible and less biased since the identification of RAD loci is done without a priori, compared to SNP chip or SSR which loci are definite and designed prior to the experiment. Compared to the 32 SSRs, the amount of data generated in a RADseq experiment would have broadened the possibility to further study the gene-flows between many more closely-related intrasectional individuals (Pante et al., 2015; Herrera and Shank, 2016; Iguchi et al., 2019).

Close taxa relationships can also be appraised through phylogeography (Avise, 2000), thus integrating environmental variables in addition to genetic diversity to infer individuals relationships. Although phylogeography is generally applied at the infraspecies level in its original definition, such approaches already provided valuable insights into the dynamics of the evolution of certain plant genera embracing closely related species, such as *Diabelia* (Zhao et al., 2019), *Helminthotheca* (Tremetsberger et al., 2016), *Pisum* (Smýkal et al., 2011) and *Zizania* (Xu et al., 2015).

5.3 A robust phylogeny for wild roses

5.3.1 Relationships between *Rosa* sections

We provided the genus *Rosa* with the most supported and comprehensive phylogenetic study ever. The evolutionary relationships between subgenera/sections are now well established. There is no reason to further consider the usual 4-subgenera scheme in *Rosa* and we advocate to treat the genus at the sectional level (Figure 40). We achieved in obtaining robust phylogenies both with plastid and nuclear sequences, especially at deep nodes which were challenging because of suspected ancient rapid radiations. We provided unprecedented insights into the phylogeny of wild roses using nuclear sequences. Although the typical two-claded structure (MC1: *Rosa*-like species, and MC2: *Synstylae*-like species) was present in both plastid and nuclear phylogenies, their relative origin is different between the two types of molecular data. The split happens at the deepest node in the plastid phylogeny while it is shallower in the nuclear phylogeny. We demonstrated that the core *Pimpinellifoliae* clade branches quite earlier, along with Hulthemia and Hesperhodos, compared to the split between MC1 and MC2 in the nuclear phylogeny. It would therefore be interesting to model a divergence time analysis on the nuclear dataset to estimate the relative ages of each bipartition. However, we are afraid that such analyses could not rely on the complete set of nuclear $\mathrm{SCO_{Tag}}$s. Indeed, dating the nuclear phylogeny would require that the dataset fit the molecular clock hypothesis. Considering that $\mathrm{SCO_{Tag}}$s have very different rates of variations, obtaining the convergence of all parameters in a Bayesian frame with a relaxed molecular clock may be intractable. On the other hand, forcing the data to fit a strict molecular clock hypothesis would be a nonsense and an easily violated hypothesis considering that $\mathrm{SCO_{Tag}}$s evolve somewhat at very different paces.

5.3.2 Considering hybridization and polyploidy processes

In addition to providing a robust phylogenetic framework, one of the major advances of our study concerns the confirmation of the hybrid origin of certain sections (Figure 38). Especially section *Caninae*, *Gallicanae* and some species of section *Pimpinellifoliae*. Although the hybrid origin of these sections were previously suspected (Roberts, 1977; Iwata et al., 2000; Fougère-Danezan et al., 2015), this was never demonstrated before in an ILS-aware, quantitative frame. Here, we identified the parental lineages of suspected hybrids and we computed the fraction inherited from each parent. The inheritance probabilities correlated well with the ploidy levels either estimated or commonly reported in the literature. Our analyses provided unprecedented insights into the reticulate evolution of the genus *Rosa*, by identifying intersectional hybrids and their parental lineages. Especially, the two main clades of *Rosa* (sect. *Rosa* and allies and sect. *Synstylae* and allies) gave rise to sect. *Caninae*, one of the most diversified section of the genus that may have expanded through rapid radiations. We provided a scenario for the appearance of *Rosa* sect. *Caninae*, and we hope that such hypothesis could be tested based on genome-wide data, possibly using $\mathrm{SCO_{Tag}}$ alleles comparisons, and compared with other hypotheses (Nybom et al., 2006; Crhak Khaitova et al., 2014; Ballmer, 2018). We demonstrated the closeness between sect. Pimpinellfoliae and sect. *Rosa*, by identifying several hybrids between those two sections. It would now be interesting to study the dynamics of these hybrid genomes compared to those from their progenitors. Are parental genomes simply juxtaposed? Are there any chromosomal rearrangements? If yes, to which extent? Again a combination between $\mathrm{SCO_{Tag}}$s targeting and long read whole genome sequencing or genomic in situ hybridization (GISH) (Younis et al., 2015) could provide valuable answer to these questions. Another advance in the understanding of *Rosa* evolution, is the study of the hybrid origin of *R. gallica* (Figure 49). Several study hypothesized that *R. gallica* originate from an intersectional hybrid between sect. *Rosa* and sect. *Synstylae* (Smulders et al., 2011; Ballmer, 2018). However, the history might be more complex. Indeed, we observed that while one of the parental lineage corresponds to sect. *Synstylae*, the other points at an ancient lineage in-between sect. *Rosa* and sect. *Synstylae*-*Caninae*. There might be three possible reasons explaining this ancient lineage, (1) we did not sample the putative parental lineage (Figure 49A), (2) it corresponds to an extinct lineage that does not exist in the wild anymore (Figure 49B), (3) *R. gallica* has a triparental origin, half from sect. *Synstylae*, a quarter from sect. *Caninae* and a quarter from sect. *Rosa* (Figure 49C). We do not think that we have missed the sampling of one major lineage since we studied most *Rosa* species except very rare, local species, so hypothesis 1 could be left out. Hypothesis 2 could be possible, given the fact that hybrid populations may represent a threat for their progenitors, with an enhanced adaptive abilities and therefore a better fitness. Sometimes, hybrid population can thrive to such an extent that they smother their parental lineages. Hypothesis 3 can be considered since sect. *Gallicanae* is somewhat more recent than the appearance of sections *Rosa*, *Caninae* and *Synstylae* (Fougère-Danezan et al., 2015). Each of those three sections could have contributed to the raise of the *Gallicanae* with possible interventions of human selection. It would therefore be interesting to consider history, genetics and environment at the same time to deal with the origin of *R. gallica* in depth.

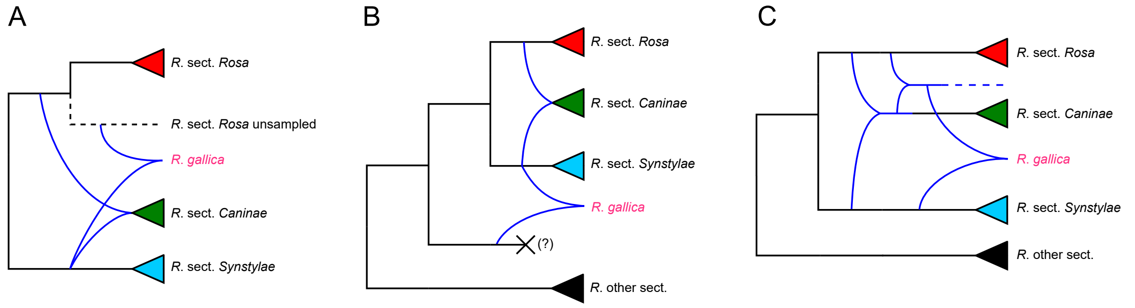


Figure 49: Three hypotheses for the origin of *Rosa gallica*. **A**. *R. gallica* originate from a cross between a *Synstylae* lineage and lineage from section *Rosa* that we did not sample in our study. **B**. *R. gallica* derive from a cross between a *Synstylae* lineage and an extinct lineage that branched at the basis of clades of sections *Synstylae* and *Rosa*. **C**. *R. gallica* has a triparental origin, resulting from a cross between a *Synstylae* lineage and a hybrid between section *Rosa* and *Caninae*.

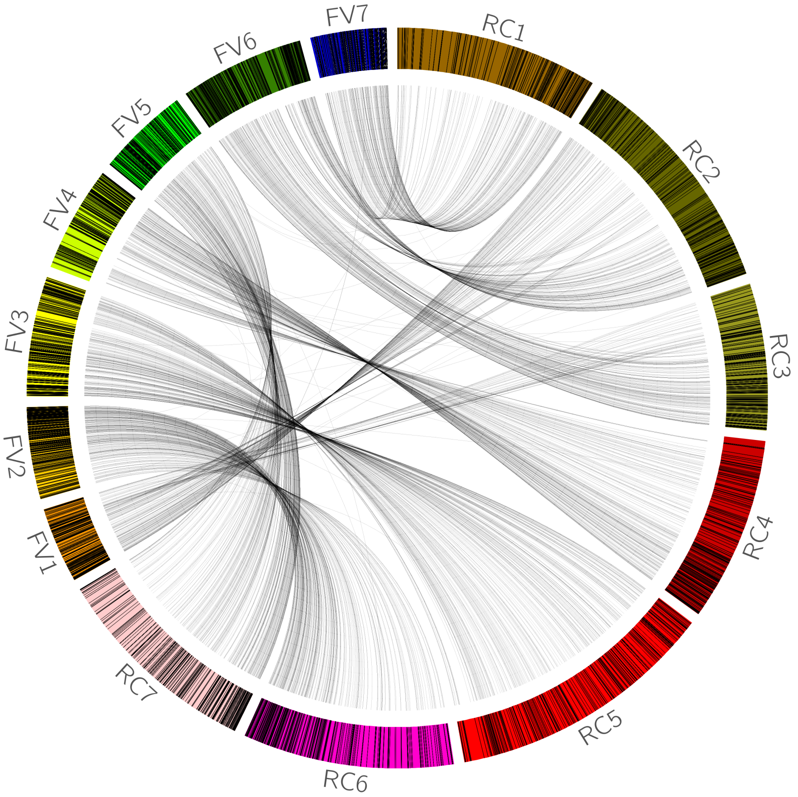
5.3.3 A framework for the study of ornamental trait evolution in roses

Roses have been widely selected by people for millennia and for many ornamental characters. Among the many traits that were selected, color, petal number, fragrance, and blooming recurrence played an important role in breeding novelties during the past centuries. Our study on *Rosa* provides a framework for studying the origin and evolution of certain traits. Gene pools in *Rosa* are generally considered from a geographical perspective, saying that European-Mediterranean roses were hardy and fragrant while Chinese roses brought recurrent blooming, some resistances (Joyaux, 2015), and a variety of new fragrances (Scalliet et al., 2008). The genetic determinisms of these traits were largely studied and several Quantitative Trait Loci (QTLs) associated with valuable ornamental characteristics were identified (Hibrand Saint-Oyant et al., 2008; Kawamura et al., 2011; Bourke et al., 2018; Smulders et al., 2019). However, few investigations were carried on the gene pool considered in the broad sense. By studying genetic diversity in genes involved in ornamental traits using a broad sampling of wild *Rosa* species, scientists and breeders may bring valuable alleles into focus. Such investigations would considerably enhance breeding programs and lead to new varieties with a series of favorable characters. This is especially true for traits associated with resistance. Pathogens are likely to bypass major resistances controlled by few alleles, but pyramiding many minor resistant alleles slows the progression of the pathogens in its host down (Pilet-Nayel et al., 2017). Indeed, host/pathogen co-evolution is compromised in the presence of a multitude of partial resistances. In a context of reduced phytochemical treatments and reluctance to the use of genetically modified organisms, at least in the European Union, the exploitation of wild resources represents more than ever an important lever for the improvement of ornamental plants. Our phylogenetic analyses represent a frame to increase the study of the wild pool of *Rosa* in prebreeding programs to identify new valuable alleles. We therefore hope to have provided a nearly exhaustive and well-structured view of the genus *Rosa*, that would then help breeders to consider less well-known species to enrich the genetic background of modern roses. By investigating aspects of hybridization and polyploidy within the genus *Rosa*, we hope to have highlighted some affinities between sections (e. g. section *Rosa* × section *Pimpinellifoliae*). This could help guide the choice of material towards compatible species in terms of hybridization affinity, phylogenetic proximity and ploidy level, thus promoting the success of introgressions from the wild gene pool. This quest for novelties comes at the time when innovation is crucial to attract consumers with new lifestyles and expectations, in the highly competitive environment of ornamental horticulture, and in a context of global climate change and chemical reduction.

Appendices

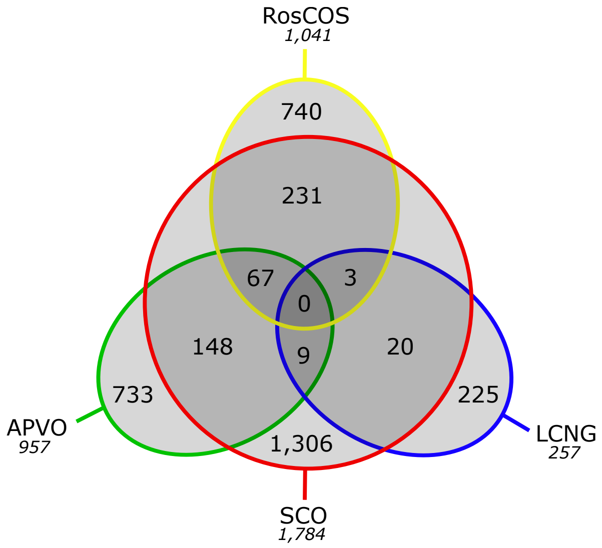
A

Supplementary tables and figures associated with Chapter 2

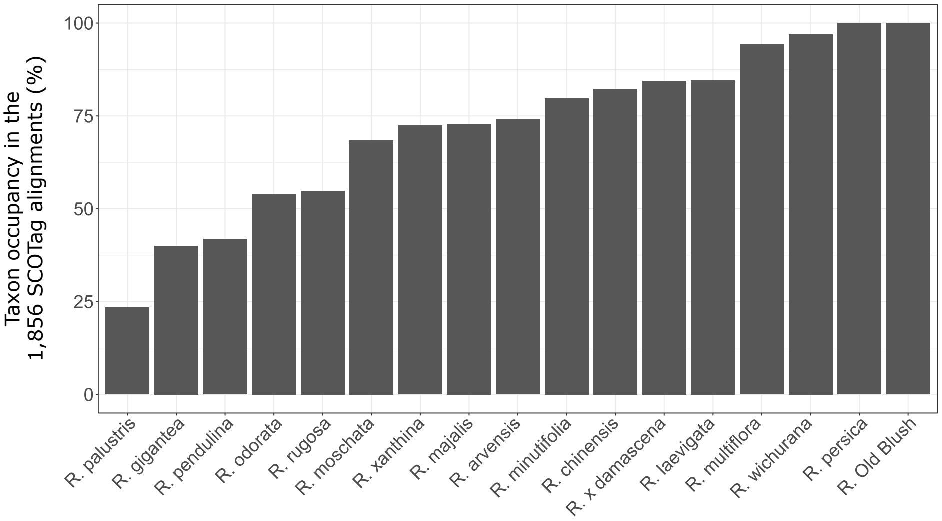


Supplementary Figure A.1: Synteny analysis of the 1784 single-copy orthologs between the genome assemblies of *Fragaria vesca* and *R.* 'Old Blush'.

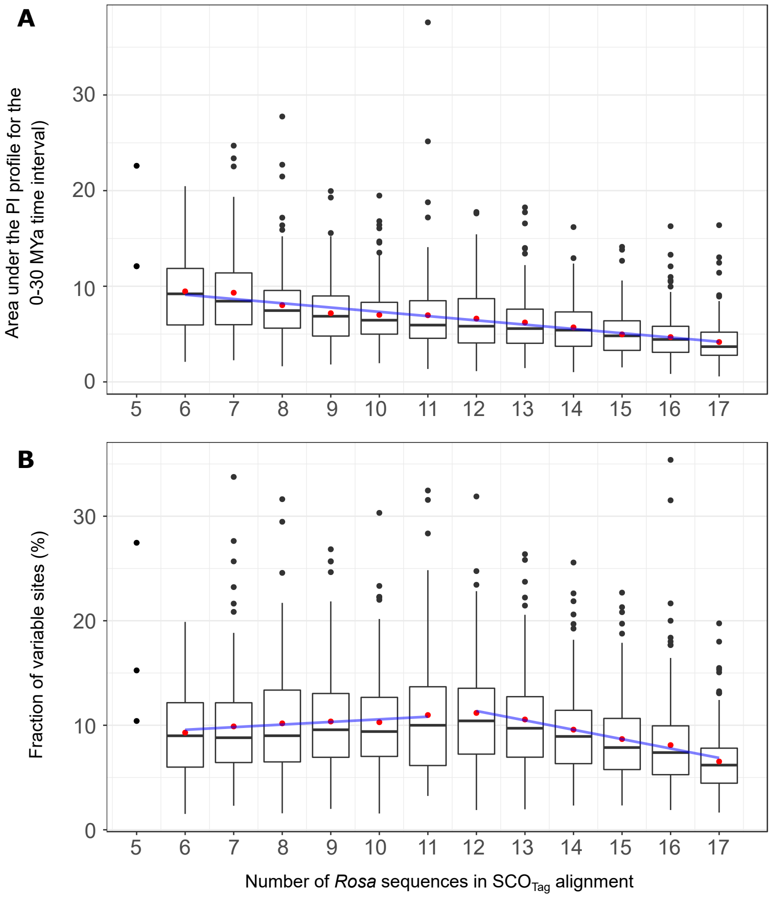
.



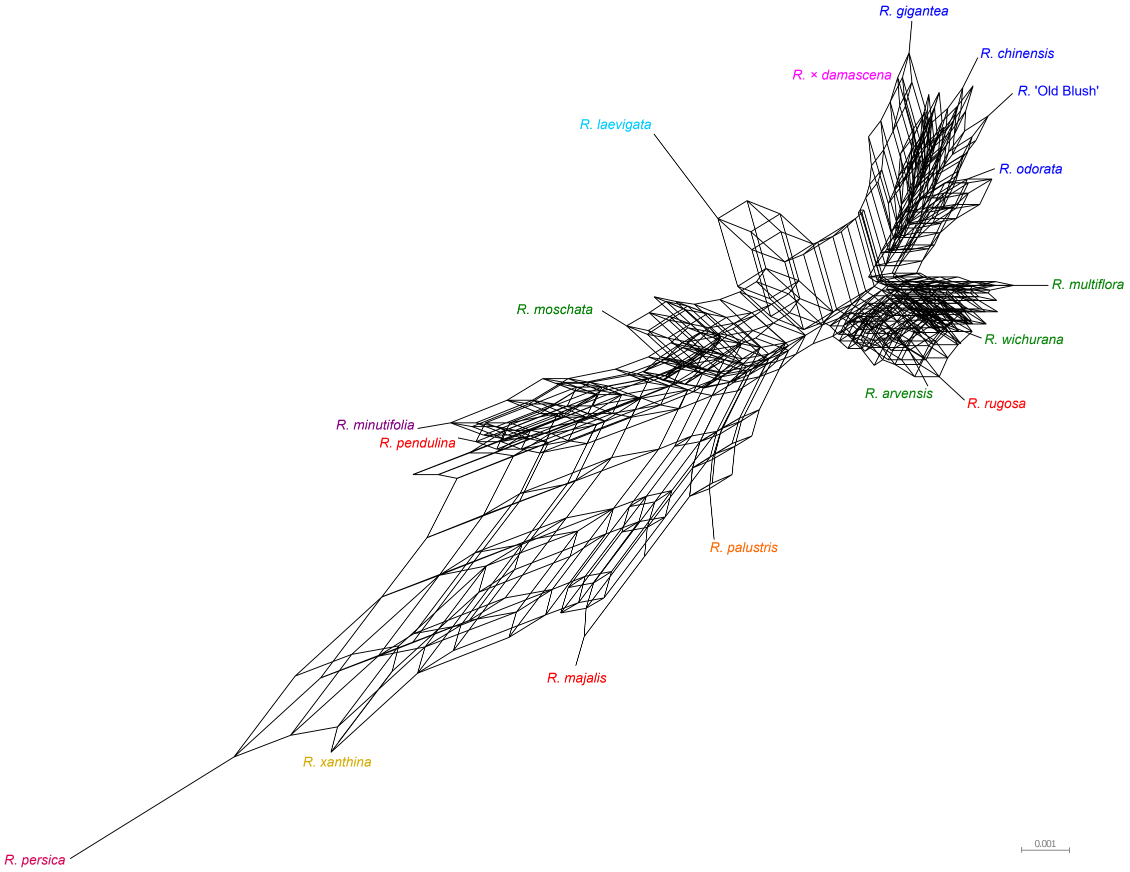
Supplementary Figure A.2: Venn diagram showing the overlapping of the 1,784 Single-Copy Ortholgs (SCO) with 3 already published datasets. APVO refers to the *Arabidopsis*, *Populus*, *Vitis*, *Oryza* dataset (Duarte et al., 2010), LCNG refers to the Low-Copy Nuclear Genes found by Liston (2014) and RosCOS corresponds to the Rosaceae Conserved Ortholog Set of markers (Cabrera et al., 2009).



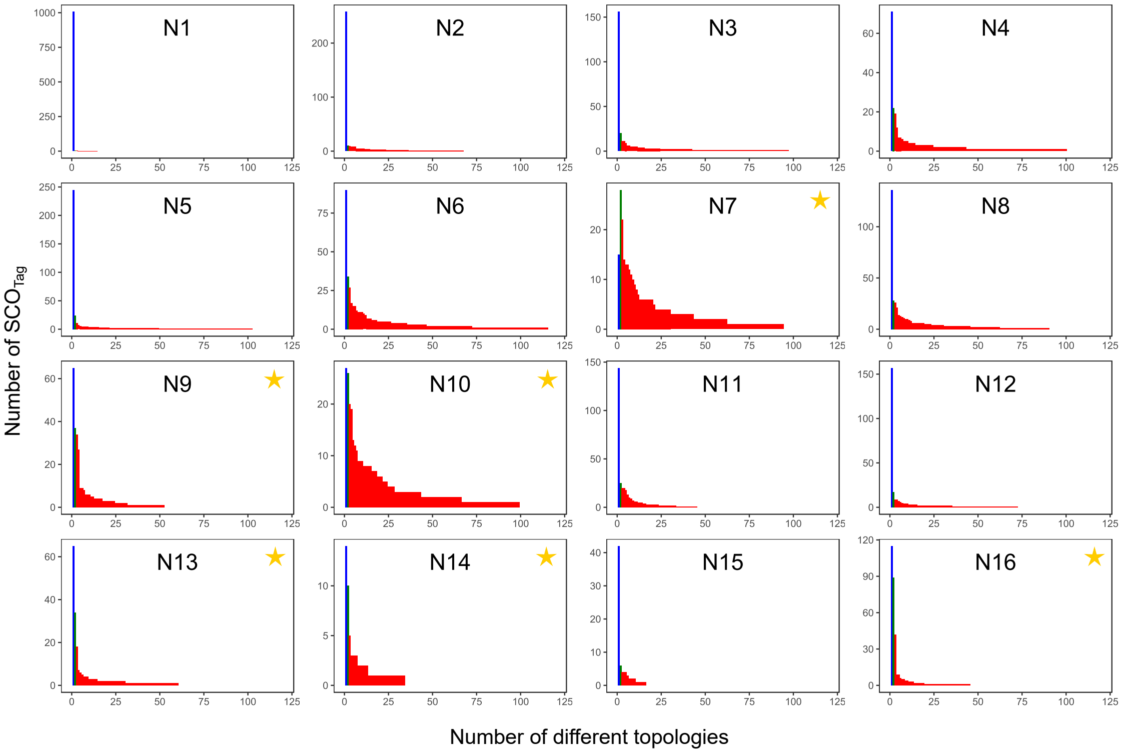
Supplementary Figure A.3: Taxon occupancy of the *Rosa* ingroup for the 1856 $\mathrm{SCO_{Tag}}$s



Supplementary Figure A.4: Impact of taxon occupancy on **A**) the area under PI profiles for the 0-30 MYa time interval ([6-17]: y=11.8-0.45x, R^2=0.18) and **B**) the fraction of variable sites in SCO_{Tag} alignment ([6-11]: $y = 8.1 + 0.25x, R^2 = 0.004$; [12-17]: $y = 22.1 - 0.90x, R^2 = 0.11$). Red dots indicate the mean values. Situations with less than 30 points were plotted but not used in the calculation of correlations.



Supplementary Figure A.5: Network representing the conflict between the 1856 SCO$_{\text{Tag}}$ trees for the *Rosa* ingroup. Species colors follow Figure 27.

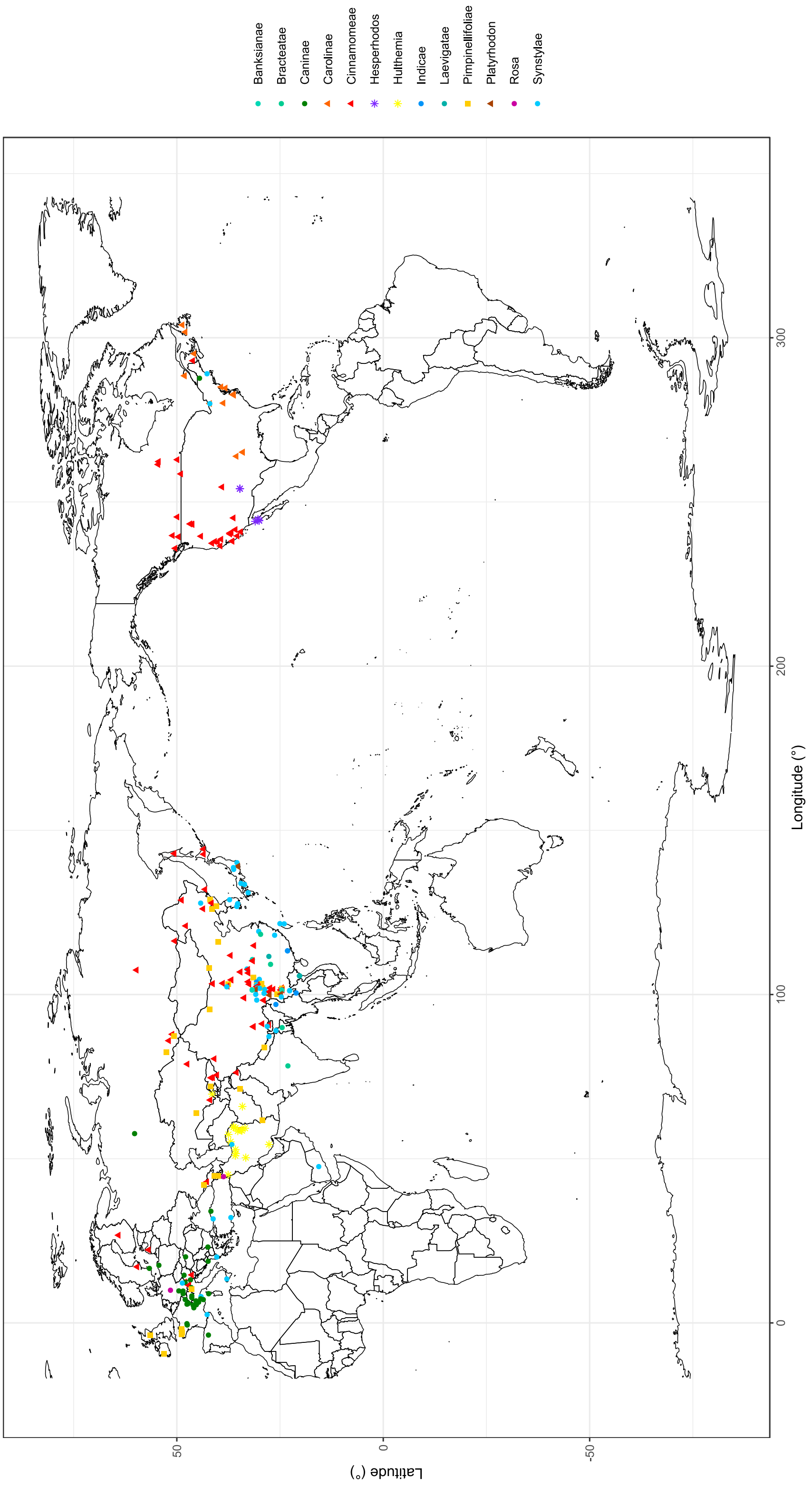


Supplementary Figure A.6: Number of SCO$_{\text{Tag}}$s supporting each bipartition. Each plot title indicates the node name as shown in Figure 4. Blue represents the SCO$_{\text{Tag}}$s that support the species-tree bipartition (ie. concordant SCO$_{\text{Tag}}$s), green indicates the SCO$_{\text{Tag}}$s that agree with the main alternative bipartition (ie. main conflicting SCO$_{\text{Tag}}$s) and red corresponds to SCO$_{\text{Tag}}$s that support other alternative bipartitions (ie. other conflicting SCO$_{\text{Tag}}$s). Stars indicate nodes with significant conflicts, ie the proportion of SCO$_{\text{Tag}}$s supporting the main alternative bipartition is greater than 50% of the proportion of SCO$_{\text{Tag}}$s supporting the species-tree bipartition.

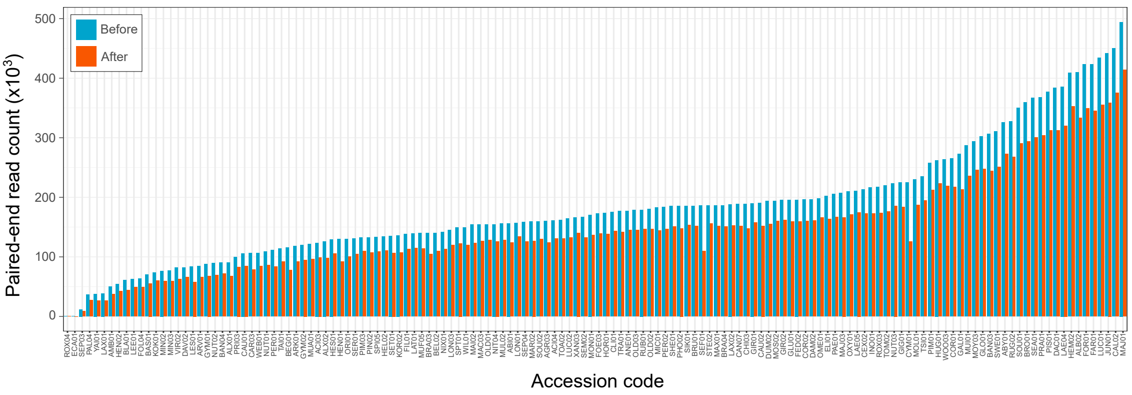
Supplementary Table A.1: Origin of plastid sequences used for phylogenetic inferences. Plastid sequences were either obtained by target-assembly/Blast or retrieved from GenBank. Dashes indicate missing data.

Species	psbA-trnH	trnG	trnL
Rosa arvensis	Recovered	Recovered	Recovered
R. chinensis	Recovered	Recovered	Recovered
R. gigantea	Recovered	Recovered	Recovered
R. laevigata	Recovered	Recovered	Recovered
R. majalis	Recovered	Recovered	Recovered
R. minutifolia	DQ778786	Recovered	Recovered
R. moschata	Recovered	Recovered	Recovered
R. multiflora	Recovered	KJ575281	KJ575162
R. 'Old Blush'	Recovered	Recovered	Recovered
R. odorata	Recovered	Recovered	Recovered
R. palustris	DQ778798	KJ575290	DQ778877
R. pendulina	Recovered	Recovered	Recovered
R. persica	Recovered	Recovered	Recovered
R. rugosa	Recovered	Recovered	Recovered
R. wichurana	Recovered	-	Recovered
R. × *damascena*	LC374596	-	KT359474
R. xanthina	Recovered	Recovered	Recovered
F. vesca	FJ493305	FJ422324	AF163559

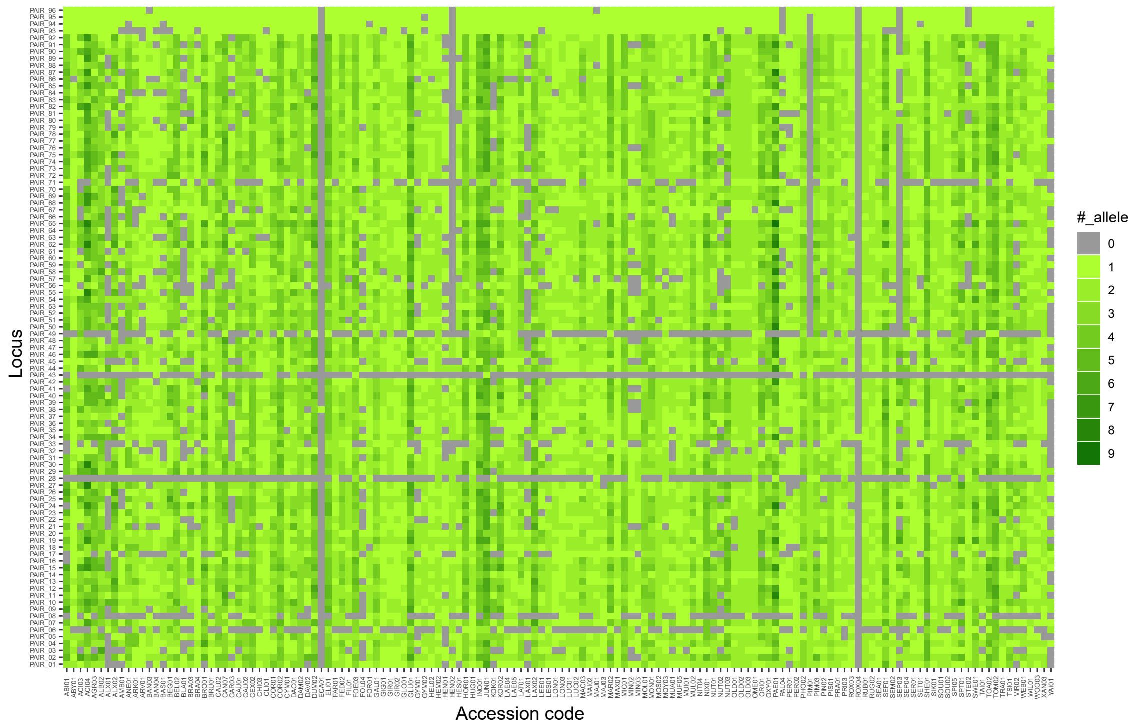
B Supplementary tables and figures associated with Chapter 3



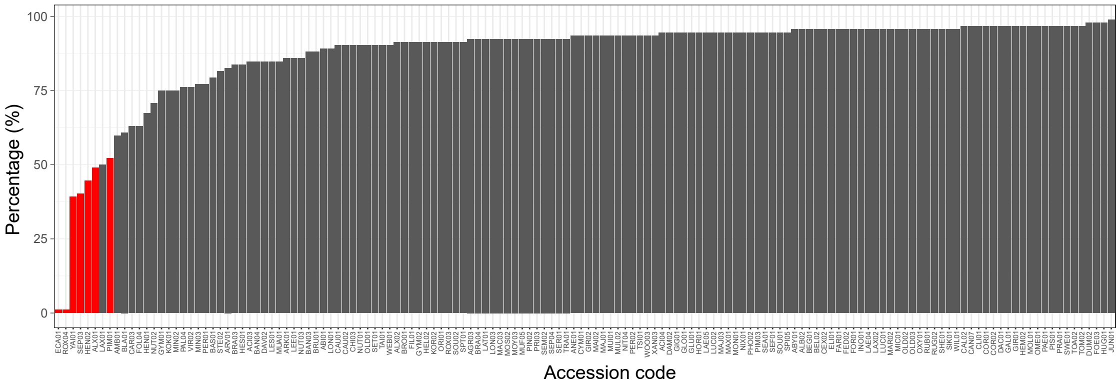
Supplementary Figure B.1: Geographical origins of the *Rosa* accessions with tissue fragments preserved at IRHS. Localizations were assigned as close as possible according to vouchers. When no precise localization was available, we attribute one region according to literature.



Supplementary Figure B.2: Paired-end read count before and after read processing. 'Before' refers to raw reads as received from the sequencing platform (end of step 1, Figure 32). 'After' refers to processed reads (step 3, Figure 32).



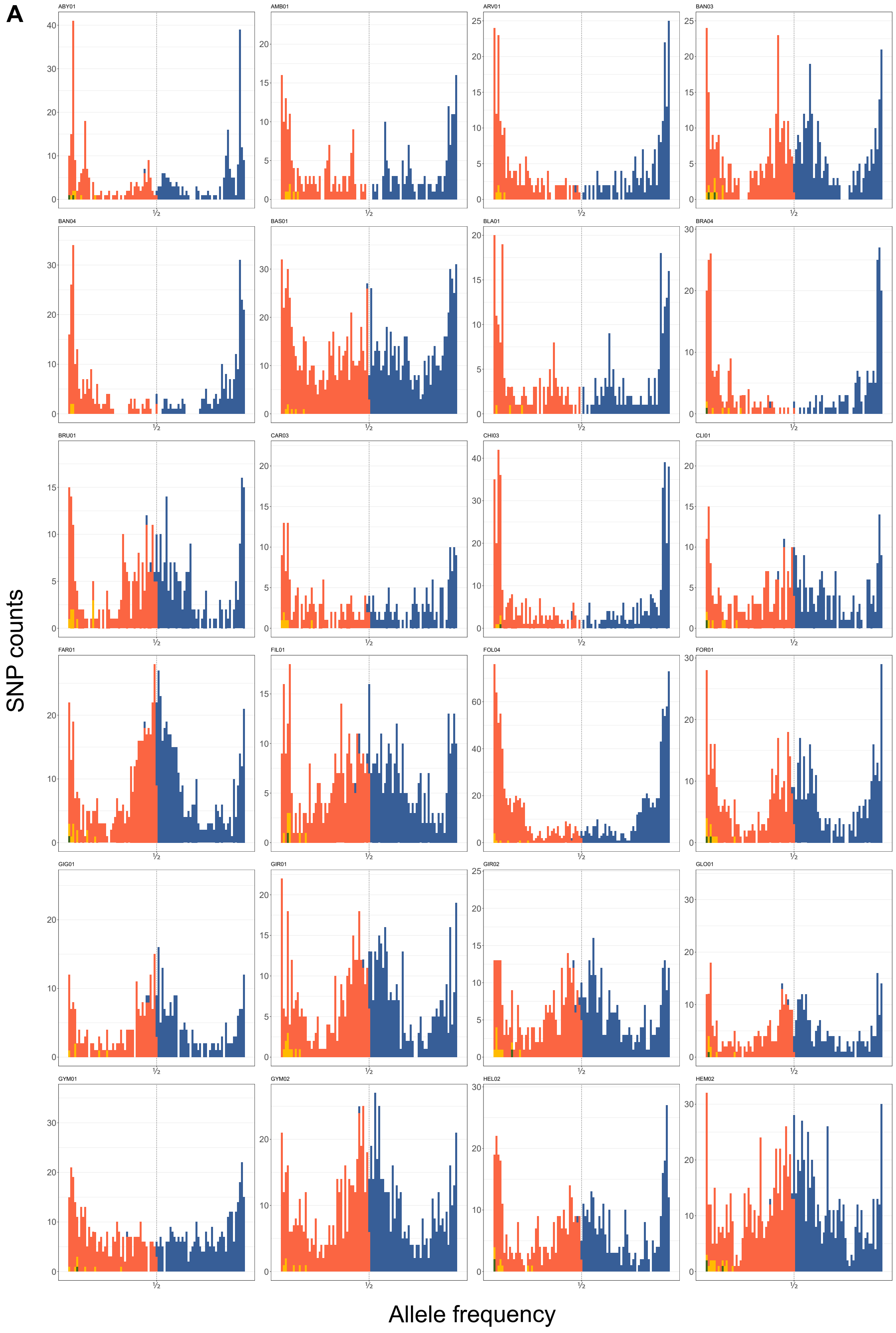
Supplementary Figure B.3: Heat map showing the number of alleles recovered for each sample at each locus.

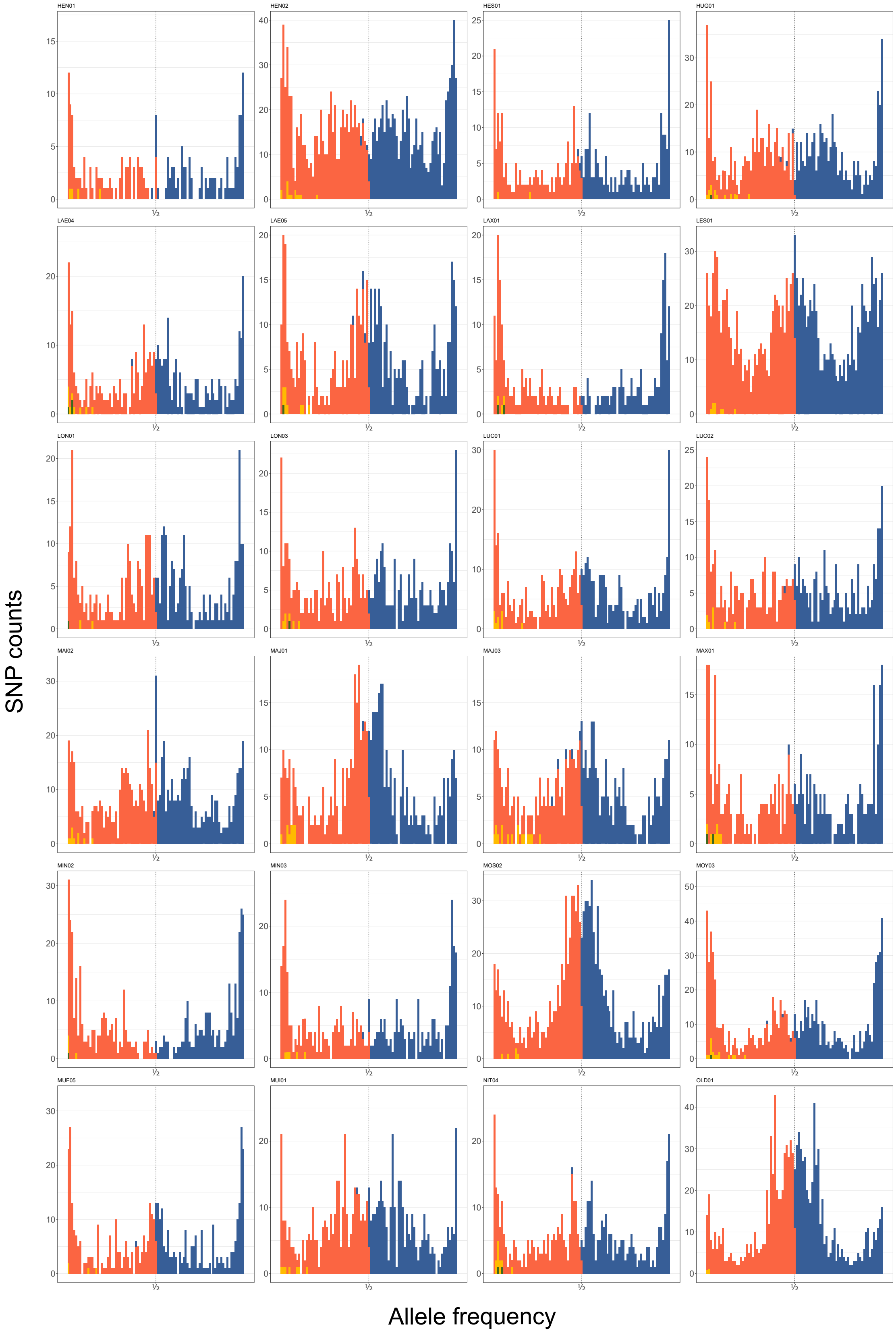


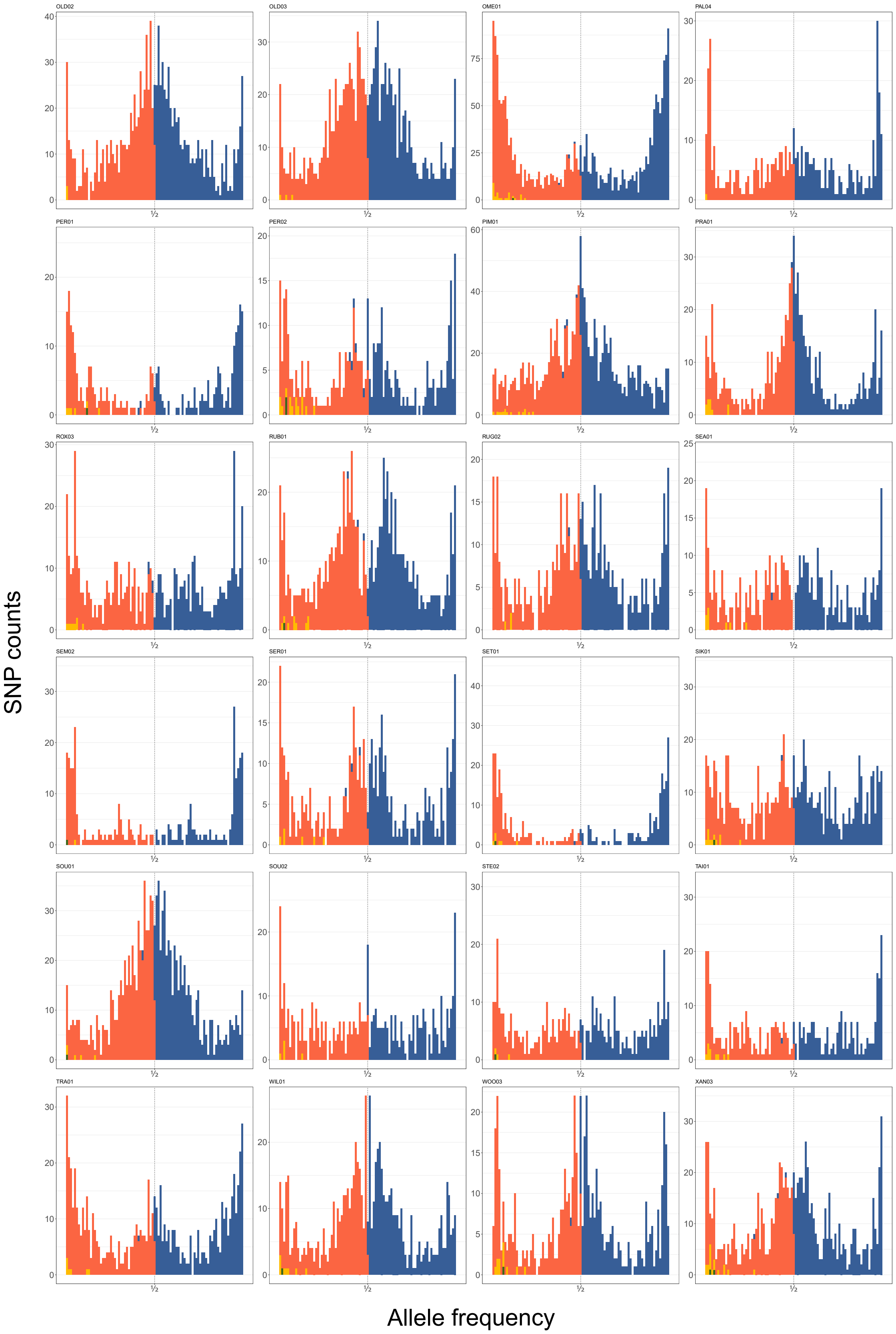
Supplementary Figure B.4: Taxon occupancy in nuclear SCO_{Tag}s. Percentages represent the fraction of nuclear SCO_{Tag}s containing at least one allele after the read processing and assembly steps. Red bars correspond to accessions with less than 50% of occupancy or with biased sequencing that were not taken into account for downstream analysis.

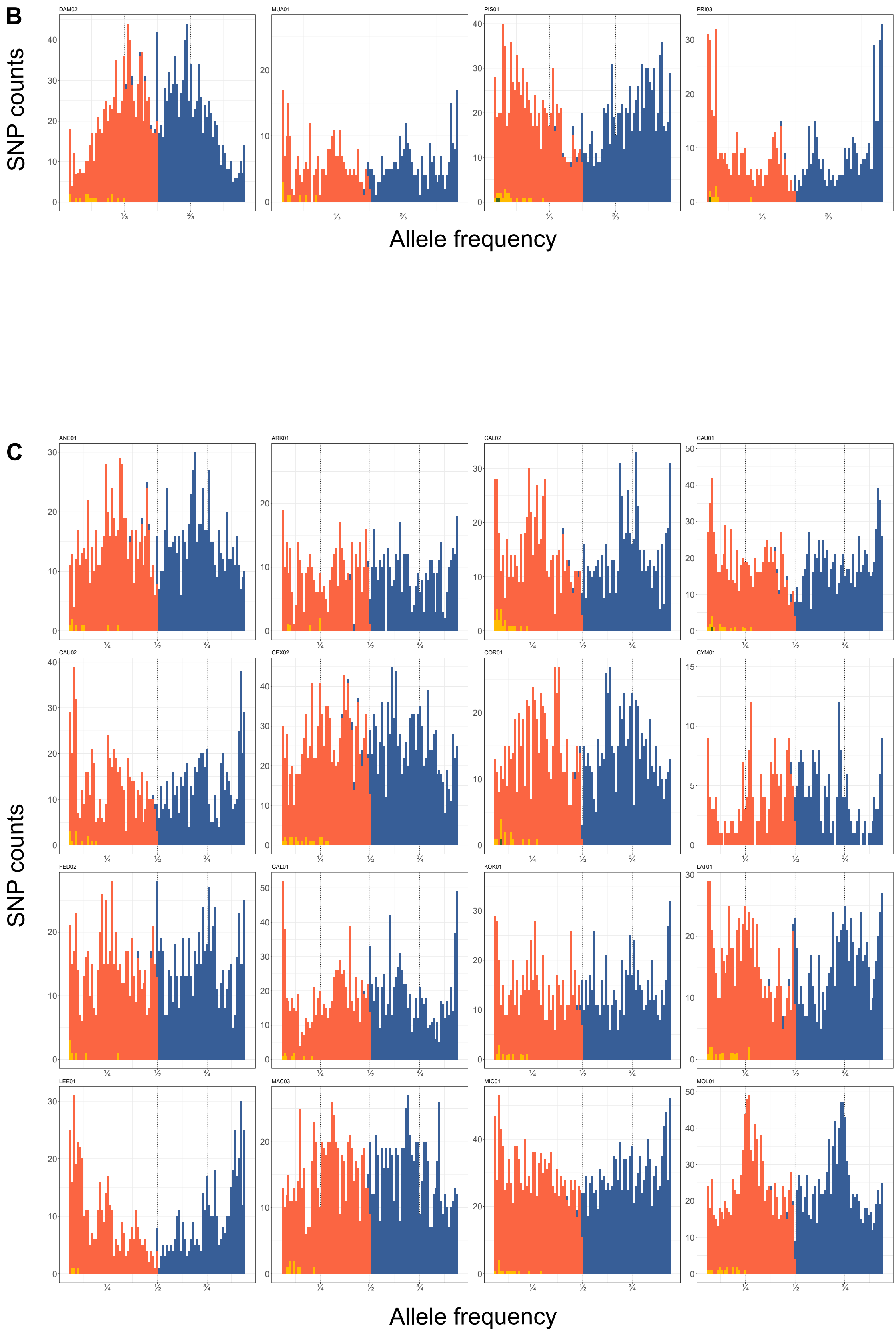
Supplementary Table B.1: Genbank references of extra *Rosa* accessions and outgroup species used for plastid phylogenies. Species for which a de novo plastid genome assembly was required have their names in bold.

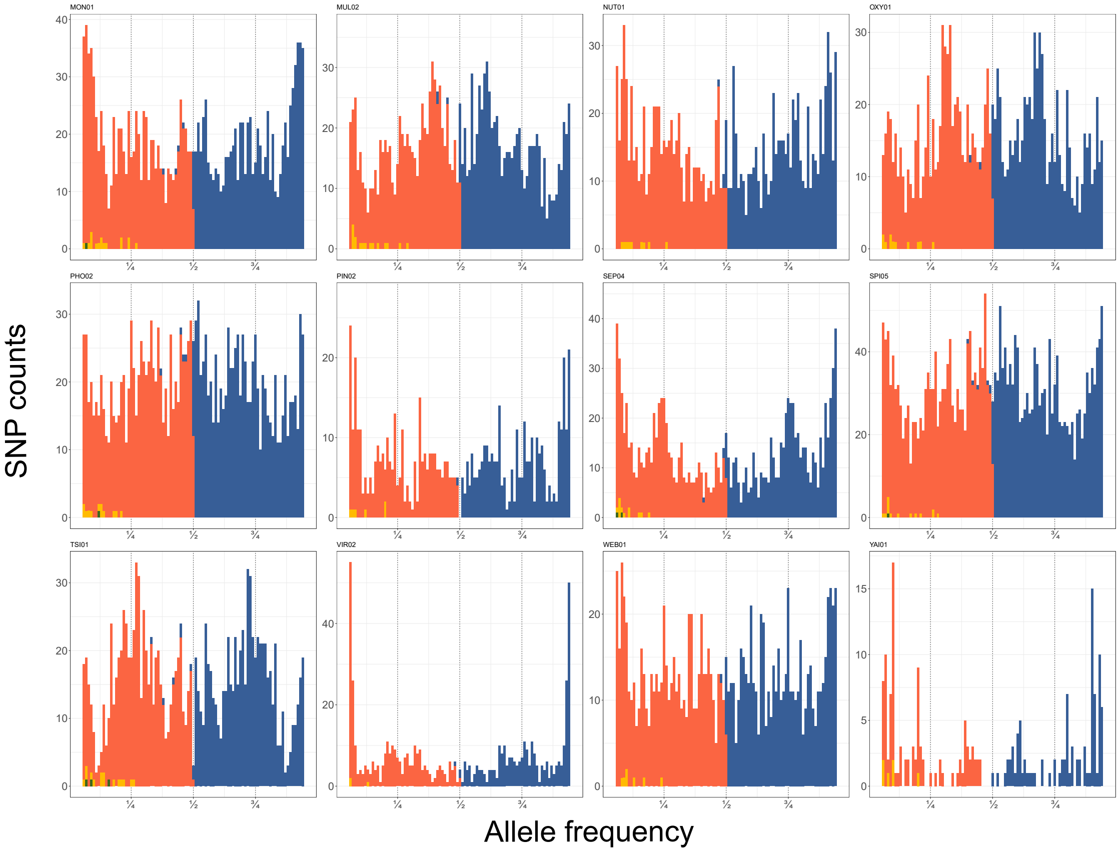
	Species	Accession code	GenBank reference
Ingroup	*Rosa arvensis*	ARV00	SRX3286288
	R. canina	CAN00	ERX1733250
	R. × *damascena*	DAM00	SRX3286290, SRX3286291
	R. dumalis	DUM00	ERX1733252
	R. elliptica	ELL00	ERX1733251
	R. gallica	GAL00	SRX4006794
	R. gigantea	GIG00	SRX3286283, SRX3286284
	R. laevigata	LAE00	SRX4006792
	R. × *odorata*	ODO00	SRX3286293
	R. 'Old Blush'	OLD00	PRJNA445774
	R. persica	PER00	SRX4006789
	R. praelucens	PAE00	MG450565
	R. roxburghii	ROX00	PRJNA356521
	R. wichurana	WIC00	SRX3286280, SRX3286281
	R. xanthina	XAN00	SRX4006788
Outgroup	*Fragaria nipponica*	na	KY769125
	Fragaria vesca	na	JF345175
	Potentilla parvifolia	na	KY420033
	Potaninia mongolica	na	KY419959
	Drymocallis glandulosa	na	KY420015
	Comarum salesovianum	na	KY420034
	Sanguisorba officinalis	na	KY419975

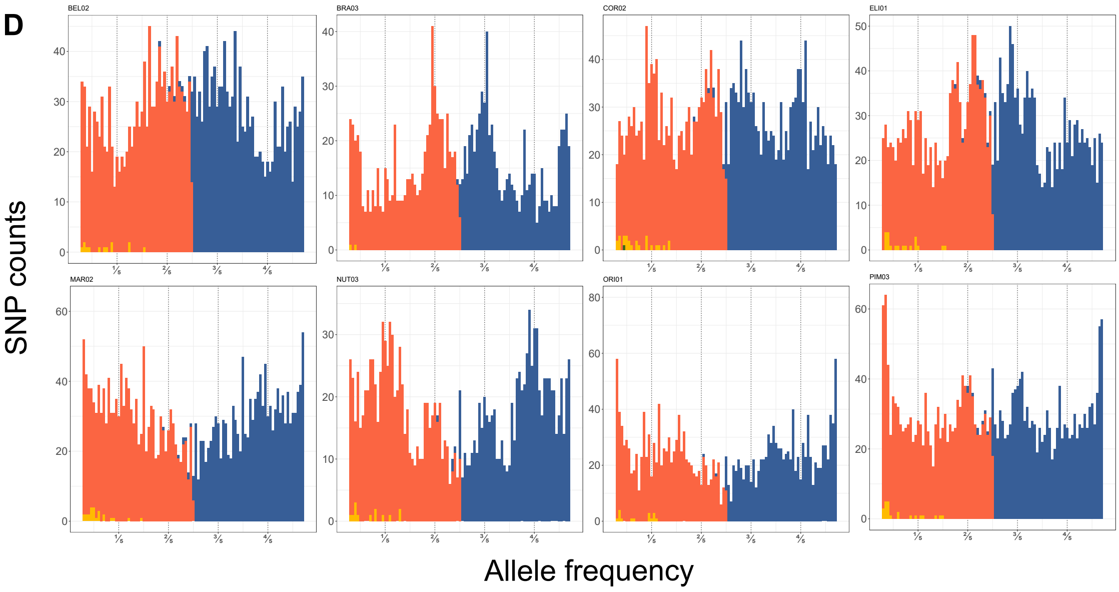

A
ABY01
AMB01
ARV01
BAN03
BAN04
BAS01
BLA01
BRA04
BRU01
CAR03
CHI03
CLI01
FAR01
FIL01
FOL04
FOR01
GIG01
GIR01
GIR02
GLO01
GYM01
GYM02
HEL02
HEM02
½
SNP counts
Allele frequency

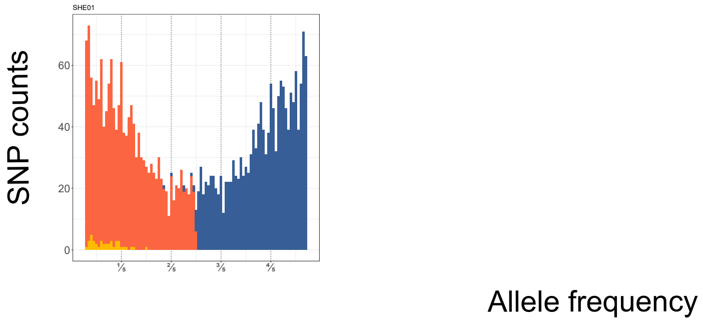

HEN01
0
5
10
15
½
HEN02
0
10
20
30
40
½
HES01
0
5
10
15
20
25
½
HUG01
0
10
20
30
½
LAE04
0
10
20
½
LAE05
0
5
10
15
20
½
LAX01
0
5
10
15
20
½
LES01
0
10
20
30
½
LON01
0
5
10
15
20
½
LON03
0
5
10
15
20
½
LUC01
0
10
20
30
½
LUC02
0
5
10
15
20
25
½
MAI02
0
10
20
30
½
MAJ01
0
5
10
15
½
MAJ03
0
5
10
15
½
MAX01
0
5
10
15
½
MIN02
0
10
20
30
½
MIN03
0
10
20
½
MOS02
0
10
20
30
½
MOY03
0
10
20
30
40
50
½
MUF05
0
10
20
30
½
MUI01
0
10
20
½
NIT04
0
10
20
½
OLD01
0
10
20
30
40
½
SNP counts
Allele frequency

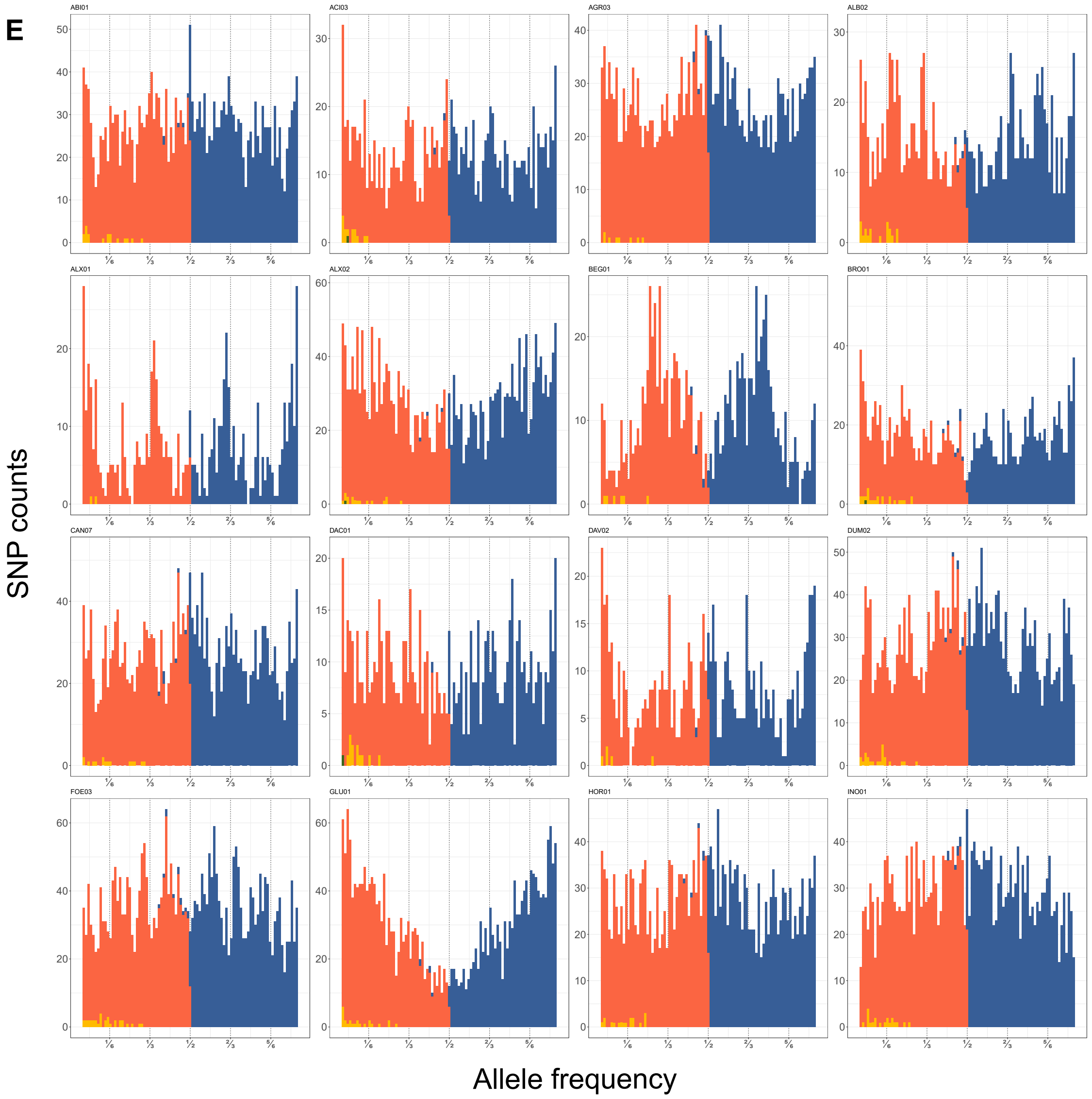

OLD02
0
10
20
30
40
½
OLD03
0
10
20
30
½
OME01
0
25
50
75
½
PAL04
0
10
20
30
½
PER01
0
10
20
½
PER02
0
5
10
15
20
½
PIM01
0
20
40
60
½
PRA01
0
10
20
30
½
ROX03
0
10
20
30
½
RUB01
0
10
20
½
RUG02
0
5
10
15
20
½
SEA01
0
5
10
15
20
25
½
SEM02
0
10
20
30
½
SER01
0
5
10
15
20
½
SET01
0
10
20
30
40
½
SIK01
0
10
20
30
½
SOU01
0
10
20
30
½
SOU02
0
10
20
½
STE02
0
10
20
30
½
TAI01
0
10
20
30
½
TRA01
0
10
20
30
½
WIL01
0
10
20
½
WOO03
0
5
10
15
20
½
XAN03
0
10
20
30
½
SNP counts
Allele frequency

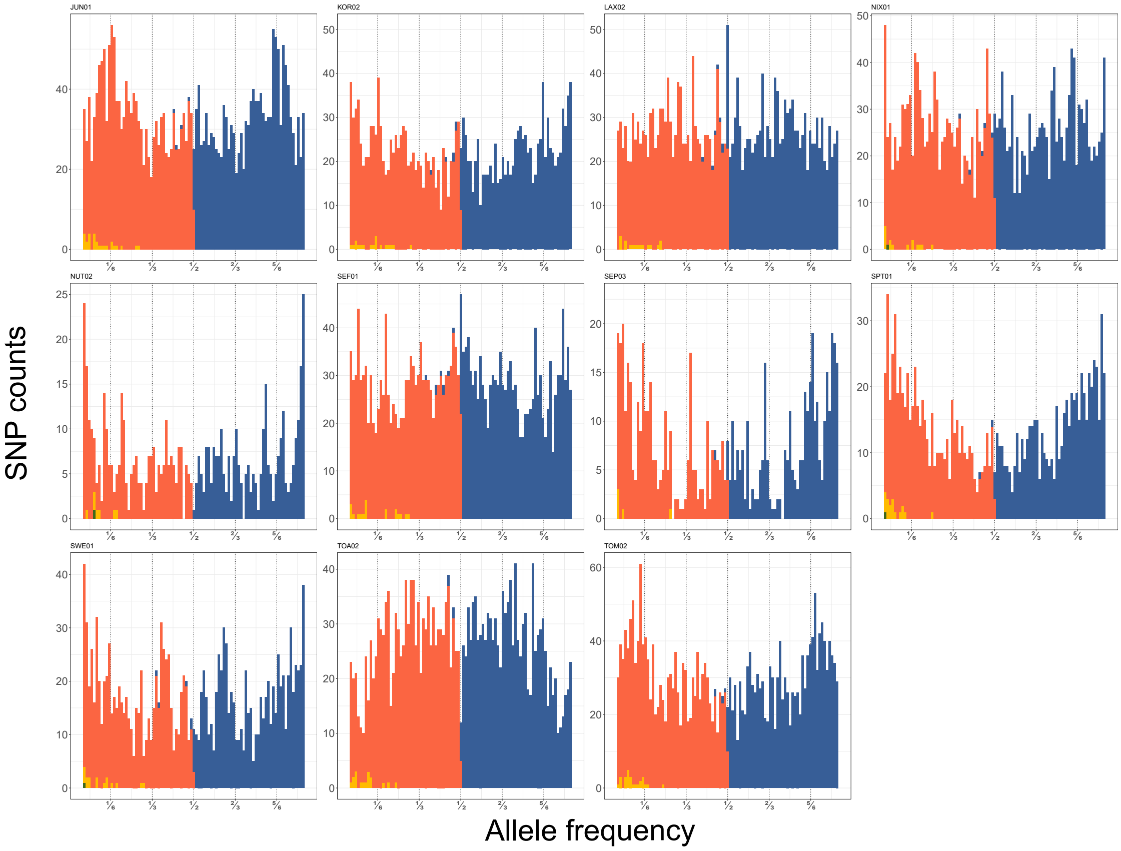

B
SNP counts
DAM02
MUA01
PIS01
PRI03
⅓
⅔
Allele frequency
C
SNP counts
ANE01
ARK01
CAL02
CAU01
CAU02
CEX02
COR01
CYM01
FED02
GAL01
KOK01
LAT01
LEE01
MAC03
MIC01
MOL01
¼
½
¾
Allele frequency


MON01
MUL02
NUT01
OXY01
PHO02
PIN02
SEP04
SPI05
TSI01
VIR02
WEB01
YAI01
SNP counts
Allele frequency
¼
½
¾

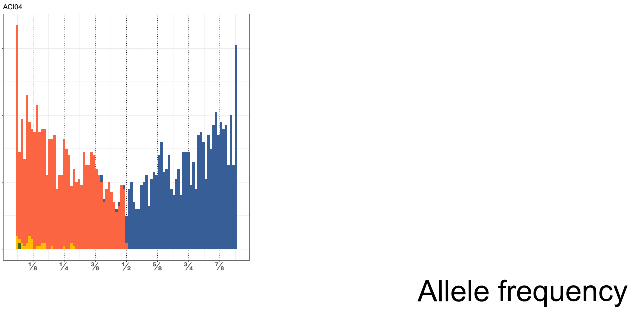

D
BEL02
BRA03
COR02
ELI01
MAR02
NUT03
ORI01
PIM03
SNP counts
Allele frequency
⅕
⅖
⅗
⅘

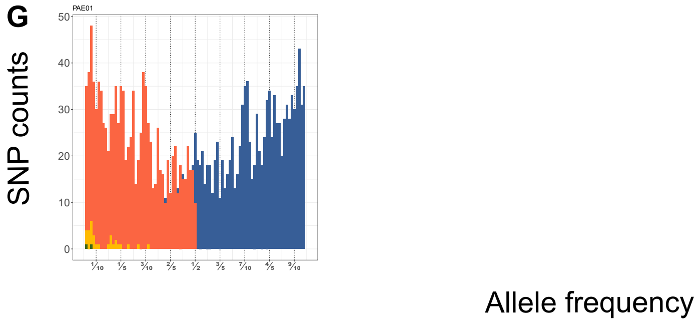

SHE01
SNP counts
0
20
40
60
1/5
2/5
3/5
4/5
Allele frequency


E
ABI01
ACI03
AGR03
ALB02
ALX01
ALX02
BEG01
BRO01
CAN07
DAC01
DAV02
DUM02
FOE03
GLU01
HOR01
INO01
SNP counts
1/6
1/3
1/2
2/3
5/6
Allele frequency

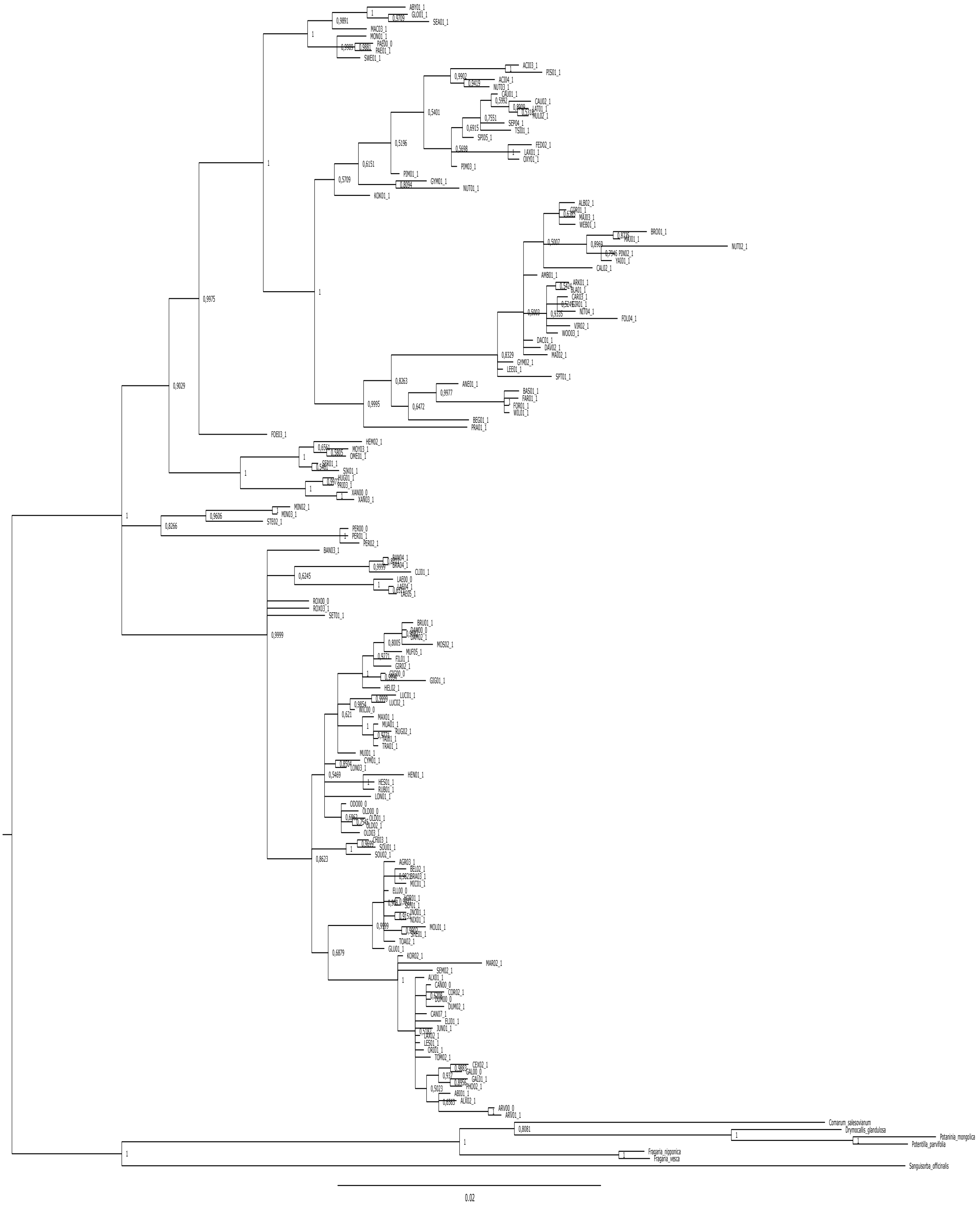

JUN01
KOR02
LAX02
NIX01
NUT02
SEF01
SEP03
SPT01
SWE01
TOA02
TOM02
SNP counts
Allele frequency
⅙
⅓
½
⅔
⅚

F

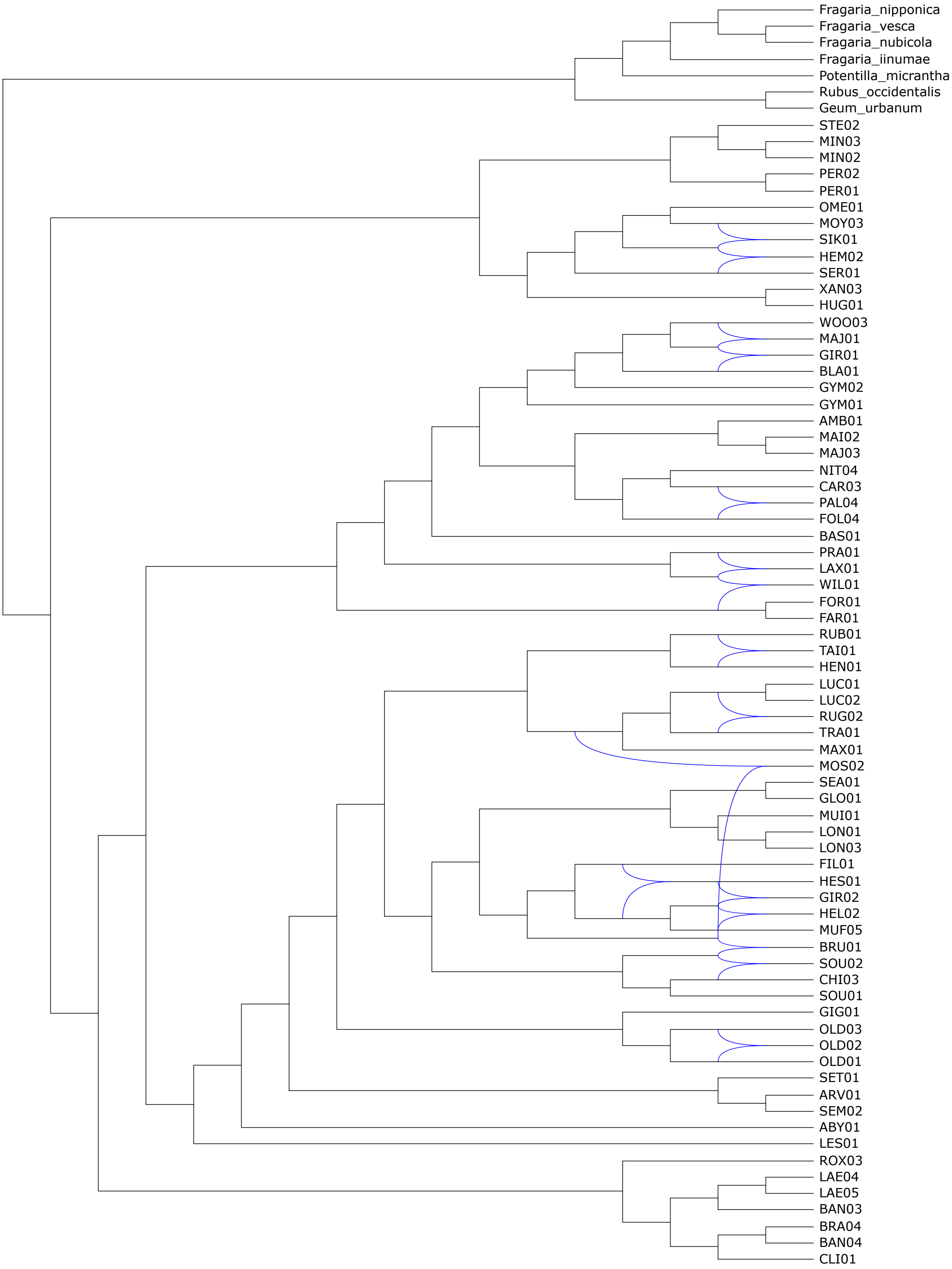

ACI04
SNP counts
Allele frequency
0
20
40
60
⅛
¼
⅜
½
⅝
¾
⅞



Supplementary Figure B.5: Estimation of the ploidy level for each accession of the study. Accessions are grouped according to their putative ploidy level as estimated by the distribution of allele frequencies with **A**) diploid, **B**) triploid, **C**) tetraploid, **D**) pentaploid, **E**) hexaploid, **F**) octoploid and **G**) decaploid. First, second, third and fourth allele at each heterozygous position is colored in blue, orange, yellow and green, respectively. Two accession (ECA01 (*R. ecae*), ROX04 (*R. roxburghii*) were not presented here due to a lack of read coverage.

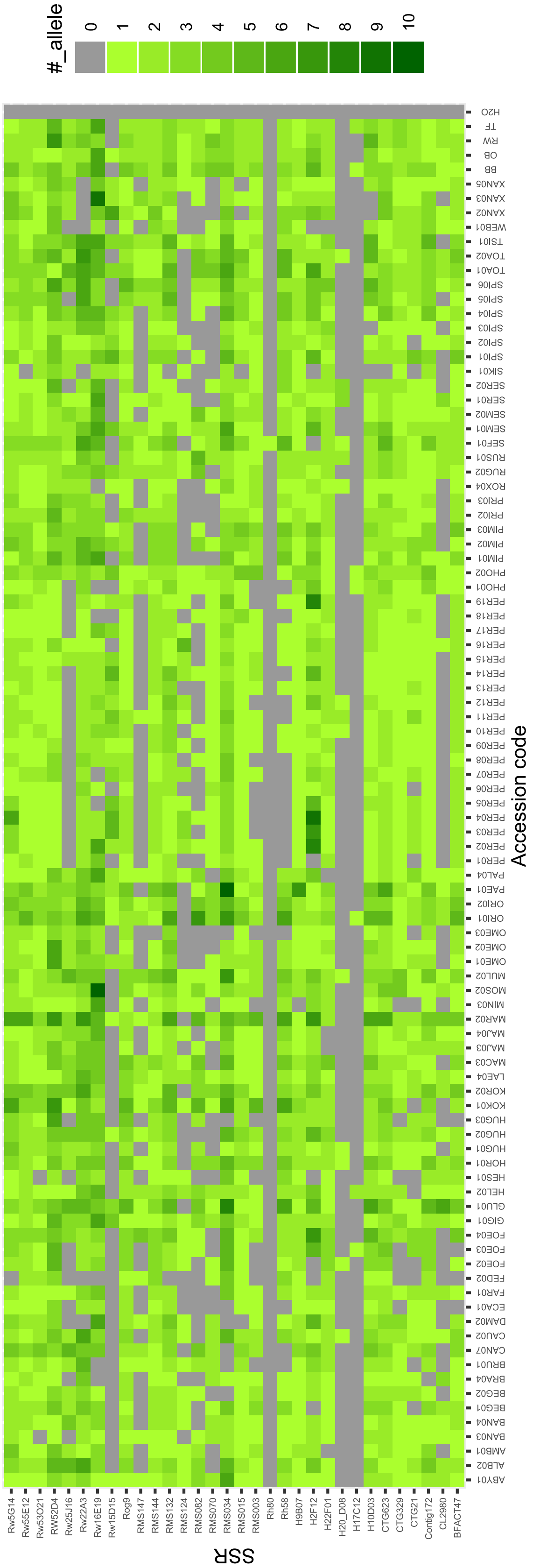


Supplementary Figure B.6: Maximum clade credibilty tree obtain after bayesian search on the concatenation of the four plastid loci. Node numbers correspond to posterior probabilities. Branches supporting bipartitions present in less than 50% of the sampled tree in the posterior distribution were collapsed. Leaf names correspond to accession codes plus the allele number.

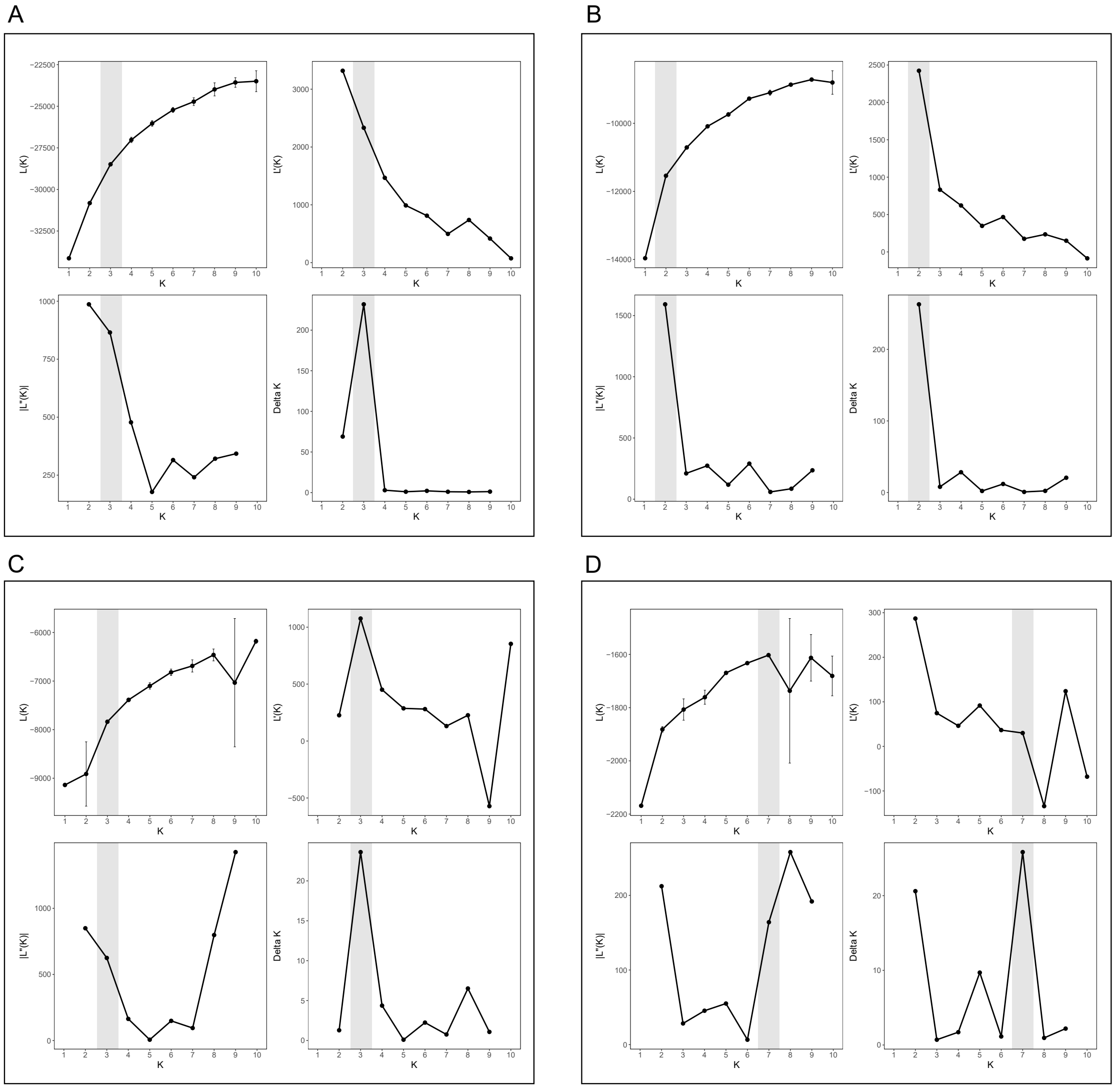


Supplementary Figure B.7: Network showing the reticulate phylogenetic relationships among diploid accessions. Nuclear allele tree were searched for a coalescent super allele tree that was further converted into a MUL-tree to obtain a hybrid network. Leaf names correspond to accession codes.

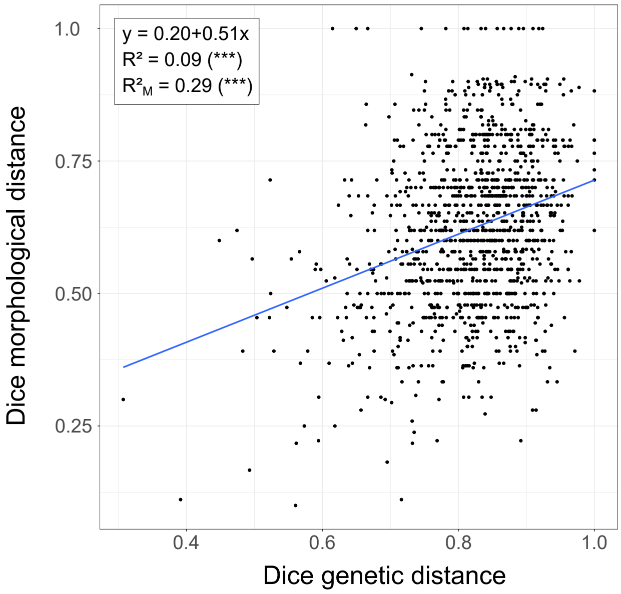
C Supplementary tables and figures associated with Chapter 4



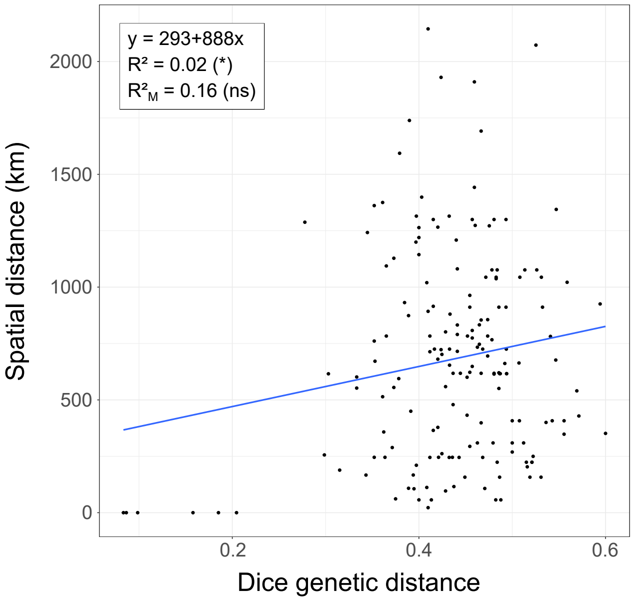
Supplementary Figure C.1: Heatmap showing the number of alleles recovered over individuals and SSR loci.



Supplementary Figure C.2: Identification of the best number of clusters (K) in structure analyses. Plots were generated in STRUCTURE Harvester following Evanno et al. (2005). The optimal number of clusters was searched for **A**. the whole genus *Rosa*, **B**. the *Pimpinellifoliae-Hulthemia* operational taxonomic unit, **C**. the *Pimpinellifoliae* section, **D**. the set of 19 accessions of *R. persica*. Each facet contains four plots. Top left hand corner: Mean likelihood L(K) and variance per K value of the data. Top right hand corner: Mean rate of change of the likelihood distribution. Bottom hand left corner: Mean absolute value of the second order rate of change of the likelihood distribution. Bottom hand right corner: Delta K = mean(|L''(K)|)/sd(L(K)). The optimal number of clusters suggested by STRUCTURE Harvester is indicated in light gray bands.



Supplementary Figure C.3: Scatter plot showing the correlation between SSR genetic diversity and morphological traits for *Rosa* samples at the genus level. Only one accession per species was considered. A linear regression was applied to model the correlation which was further tested using the Mantel's test (*** corresponds to p-value $< 1e^{-16}$).



Supplementary Figure C.4: Scatter plot showing the correlation between SSR genetic diversity and spatial distance for 19 samples of *Rosa persica*. The pairwise spatial distances between samples were computed given the latitude and longitude coordinates of each sample. A linear regression was applied to model the correlation which was further tested using the Mantel's test. (* corresponds to p-value < 0.05; ns: for non-significant).

Supplementary Table C.1: Accessions used in the study. The accession code follows Chapter 3. Nomenclature corresponds to Masure (2013). NA mean that origin record was not available.

Accession code	Section	Species	Locality
ABY01	*Synstylae*	*R. abyssinica* Lindl.	Yemen
ALB02	*Cinnamomeae*	*R. albertii* Regel	Russia
AMB01	*Cinnamomeae*	*R. amblyotis* CA Meyer	Hokkaido, Japan
BAN03	*Banksianae*	*R. banksiae* Ait.	Hubei, China
BAN04	*Banksianae*	*R. banksiae* Ait.	Yunnan, China
BEG01	*Cinnamomeae*	*R. beggeriana* Schrenk	NA
BEG02	*Cinnamomeae*	*R. beggeriana* Schrenk	Kyrgyzstan
BRA04	*Bracteatae*	*R. bracteata* Wendl.	NA
BRU01	*Synstylae*	*R. brunonii* Lindl.	NA
CAN07	*Caninae*	*R. canina* L.	Doubs, France
CAU02	*Cinnamomeae*	*R. caudata* Baker	Gansu, China
DAM02	*Rosa*	*R. damascena* L.	NA
ECA01	*Pimpinellifoliae*	*R. ecae* Aitch.	NA
FAR01	*Pimpinellifoliae*	*R. farreri* Stapf ex Cox	NA
FED02	*Cinnamomeae*	*R. fedtschenkoana* Regl.	NA
FOE02	*Pimpinellifoliae*	*R. foetida* Herrm.	NA
FOE03	*Pimpinellifoliae*	*R. foetida* Herrm.	NA
FOE04	*Pimpinellifoliae*	*R. foetida* Herrm. var. *bicolor* (Jacq.) Willm.	NA
GIG01	*Indicae*	*R. gigantea* Collett ex. Crép.	Yunnan, China
GLU01	*Caninae*	*R. glutinosa* Sibth. & Sm. var. *dalmatica* (A.Kern.) C.K.Schneid.	Montenegro
HEL02	*Synstylae*	*R. helenae* Rehder & E.H.Wilson	Gansu, China
HES01	*Pimpinellifoliae*	*R. hemisphaerica* var. *rapinii* (Boiss. & Bal.) Rowlee	NA
HOR01	*Caninae*	*R. horrida* Fisch.	NA
HUG01	*Pimpinellifoliae*	*R. hugonis* Hemsl.	NA
HUG02	*Pimpinellifoliae*	*R. hugonis* Hemsl.	NA
HUG03	*Pimpinellifoliae*	*R. hugonis* Hemsl.	NA
KOK01	*Pimpinellifoliae*	*R. kokanica* Reg. Ex Juz.	Kyrgystan
KOR02	*Pimpinellifoliae*	*R. koreana* Komarov	Korea
LAE04	*Laevigatae*	*R. laevigata* Michx.	NA
MAC03	*Cinnamomeae*	*R. macrophylla* Lindl.	Bhutan
MAJ03	*Cinnamomeae*	*R. majalis* Herrm.	Russia
MAJ04	*Cinnamomeae*	*R. majalis* Herrm.	Finland
MAR02	*Caninae*	*R. marginata* Wallr.	Baden-Württemberg, Germany
MIN03	*Hesperhodos*	*R. minutifolia* Engelm.	Baja California, Mexico
MOS02	*Synstylae*	*R. moschata* Herrm.	Bhutan
MUL02	*Cinnamomeae*	*R. multibracteata* Hemsl. & Wils.	Sichuan, China
OME01	*Pimpinellifoliae*	*R. omeiensis* Rolfe	Sichuan, China
OME02	*Pimpinellifoliae*	*R. omeiensis* f. *chrysocarpa* Redh.	NA
OME03	*Pimpinellifoliae*	*R. omeiensis* var *omeiensis* f. *pteracantha* (Franch) Rehd. & Wils.	NA
ORI01	*Caninae*	*R. orientalis* Dupont ex Seringe	Russia
ORI02	*Caninae*	*R. orientalis* Dupont ex Ser.	NA
PAE01	*Platyrhodon*	*R. praelucens* Byhouwer	Yunnan, China
PAL04	*Carolinae*	*R. palustris* Marshall	New Jersey, USA
PER01	*Hulthemia*	*R. persica* Michx.	Tashkent, Uzbekistan
PER02	*Hulthemia*	*R. persica* Michx.	Mashhad, Iran
PER03	*Hulthemia*	*R. persica* Michx.	Mashhad, Iran
PER04	*Hulthemia*	*R. persica* Michx.	Mashhad, Iran
PER05	*Hulthemia*	*R. persica* Michx.	Mashhad, Iran
PER06	*Hulthemia*	*R. persica* Michx.	Gonabad, Iran
PER07	*Hulthemia*	*R. persica* Michx.	Firoozkooh, Iran
PER08	*Hulthemia*	*R. persica* Michx.	Sorkhehesar, Iran
PER09	*Hulthemia*	*R. persica* Michx.	Urmia, Iran
PER10	*Hulthemia*	*R. persica* Michx.	Birjand, Iran
PER11	*Hulthemia*	*R. persica* Michx.	Moghan, Iran
PER12	*Hulthemia*	*R. persica* Michx.	Bojnord, Iran
PER13	*Hulthemia*	*R. persica* Michx.	Ferdows, Iran
PER14	*Hulthemia*	*R. persica* Michx.	Robatsefid, Iran
PER15	*Hulthemia*	*R. persica* Michx.	Karaj, Iran
PER16	*Hulthemia*	*R. persica* Michx.	Khansar, Iran
PER17	*Hulthemia*	*R. persica* Michx.	Lar, Iran
PER18	*Hulthemia*	*R. persica* Michx.	Golestan, Iran
PER19	*Hulthemia*	*R. persica* Michx.	Afghanistan
PHO01	*Synstylae*	*R. phoenicea* Boiss.	NA
PHO02	*Synstylae*	*R. phoenicea* Boiss.	Zonguldak, Turkey
PIM01	*Pimpinellifoliae*	*R. pimpinellifolia* L.	Scotland, UK
PIM02	*Pimpinellifoliae*	*R. pimpinellifolia* L.	Brittany, France
PIM03	*Pimpinellifoliae*	*R. pimpinellifolia* L.	Normandy, France
PRI02	*Pimpinellifoliae*	*R. primula* Boulenger	NA
PRI03	*Pimpinellifoliae*	*R. primula* Boulenger	NA
ROX04	*Platyrhodon*	*R. roxburghii* Tratt.	China
RUG02	*Cinnamomeae*	*R. rugosa* Thunb.	Hokkaido, Japan
RUS01	*Synstylae*	*R. ruscinonensis* Gren. & Déségl. ex Déségl.	Pyrénées-Orientales, France
SEF01	*Caninae*	*R. serafinii* Viv.	Corsica, France
SEM01	*Synstylae*	*R. sempervirens* L.	NA
SEM02	*Synstylae*	*R. sempervirens* L.	Palermo, Italy
SER01	*Pimpinellifoliae*	*R. sericea* Lindl.	Nepal
SER02	*Pimpinellifoliae*	*R. sericea* Lindl.	Yunnan, China
SIK01	*Pimpinellifoliae*	*R. sikangensis* T.T.Yu & T.C.Ku	Yunnan, China
SPI01	*Pimpinellifoliae*	*R. spinosissima* L.	NA
SPI02	*Pimpinellifoliae*	*R. spinosissima* L.	Karachay-Cherkessia, Russia
SPI03	*Pimpinellifoliae*	*R. spinosissima* L.	Altajski Kraj, Russia
SPI04	*Pimpinellifoliae*	*R. spinosissima* L.	Clare county, Ireland
SPI05	*Pimpinellifoliae*	*R. spinosissima* L.	Bretagne, France
SPI06	*Pimpinellifoliae*	*R. spinosissima* L. var. *altaica* (Willd.) Rehd.	NA
TOA01	*Caninae*	*R. tomentosa* Sm.	NA
TOA02	*Caninae*	*R. tomentosa* Sm.	Haute-Saône, France
TSI01	*Pimpinellifoliae*	*R. tsinglingensis* Pax. & Hoffm.	Gansu, China
WEB01	*Cinnamomeae*	*R. webbiana* Royle	NA
XAN02	*Pimpinellifoliae*	*R. xanthina* var. *allardii* hort.	NA
XAN03	*Pimpinellifoliae*	*R. xanthina* Lindl.	NA
XAN05	*Pimpinellifoliae*	*R. xanthina* Lindl.	NA

Supplementary Table C.2: Summary of the new alleles found at each SSR locus. Alleles were considered new if they did not belong to the bin set developed in Liorzou et al. (2016).

SSR ID	New alleles	#	SSR ID	New alleles	#
Rog9	na	0	Rw53O21	na	0
Rw52D24	149	1	Rh80	na	0
RMS070B	261-213-260-192-240-232-238-224-259-144-230-317-109-207-251-345	16	H22F01	249-245-273-267	4
RMS124	156-170	2	RMS144	207-192-117-224	4
Rw22A3	165	1	CTGROW21	201-126	2
RMS082	147-154-185-240-301-204-177-215	8	RMS132	135-244-109-343-194-155	6
RMS015	166-150-172-174	4	Rw5G14	245-258-157-323-287-332-159-329	8
H10D03	248-203-211-198-302-258-391-170-348	9	H2F12	193-269-162-175-136-102-266-251-120-140-165-327	12
H20D08	224-179-164-199	4	CTGROW623	380-201-351-197-213-252-131-260-206-99	10
Contig172	151-142-199	3	Rw55E12	109-163-148	3
H9B07	244-230-209	3	RMS140	155-137-131-303	4
Rw16E19	324-111-142-382-136-154-269-237-283-114-308-265-112-157-351-235-329-377-246-186-207-166-160-315-286-369-183	27	H17C12	144-145-150	3
RMS034	168-111-144-164-170-112-103-159	8	RMS003	172-337-233	3
CLROW2980	145-144-237-215-126-140-334	7	CTGROW329	217	1
Rw15D15	117-147-150-133-136-142-144-139-129-146-393-157	12	Rw25J16	137-136-113-143-117-126	6
Rh58	257-214-278-283	4	BFACT47	138-135-139	3

Bibliography

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J. E., Bierne, N., Boughman, J., Brelsford, A., Buerkle, C. A., Buggs, R., Butlin, R. K., Dieckmann, U., Eroukhmanoff, F., Grill, A., Cahan, S. H., Hermansen, J. S., Hewitt, G., Hudson, A. G., Jiggins, C., Jones, J., Keller, B., Marczewski, T., Mallet, J., Martinez-Rodriguez, P., Möst, M., Mullen, S., Nichols, R., Nolte, A. W., Parisod, C., Pfennig, K., Rice, A. M., Ritchie, M. G., Seifert, B., Smadja, C. M., Stelkens, R., Szymura, J. M., Väinölä, R., Wolf, J. B. W., and Zinner, D. (2013). Hybridization and speciation. *Journal of Evolutionary Biology*, 26(2):229–246.

Ahmed, I. (2015). Chloroplast Genome Sequencing: Some Reflections. *Journal of Next Generation Sequencing & Applications*, 2(2).

Al-Nakeeb, K., Petersen, T. N., and Sicheritz-Pontén, T. (2017). Norgal: extraction and de novo assembly of mitochondrial DNA from whole-genome sequencing data. *BMC Bioinformatics*, 18(510).

Alboukadel, K. and Mundt, F. (2017). factoextra: Extract and Visualize the Results of Multivariate Data Analyses.

Alfaro, M. E., Zoller, S., and Lutzoni, F. (2003). Bayes or Bootstrap? A Simulation Study Comparing the Performance of Bayesian Markov Chain Monte Carlo Sampling and Bootstrapping in Assessing Phylogenetic Confidence. *Molecular Biology and Evolution*, 20(2):255–266.

Alix, K., Gérard, P. R., Schwarzacher, T., and Heslop-Harrison, J. S. P. (2017). Polyploidy and interspecific hybridization: partners for adaptation, speciation and evolution in plants. *Annals of Botany*, 120(2):183–194.

Allen, J. M., Huang, D. I., Cronk, Q. C., and Johnson, K. P. (2015). aTRAM - automated target restricted assembly method: a fast method for assembling loci across divergent taxa from next-generation sequencing data. *BMC Bioinformatics*, 16(98).

Allen, J. M., LaFrance, R., Folk, R. A., Johnson, K. P., and Guralnick, R. P. (2018). aTRAM 2.0: An Improved, Flexible Locus Assembler for NGS Data. *Evolutionary Bioinformatics Online*, 14.

Amrine, J. W. (2002). Multiflora rose. In *Biological Control of Invasive Plants in the Eastern United States*, page 413. Van Driesche R., USDA Forest Service Publication FHTET edition.

Andersen, H. L., Næss, S. J., and Salvesen, P. H. (2016). Hybridization between the locally endangered *Rosa spinosissima* and *Rosa mollis* results in the pentaploid *Rosa* × *sabinii* in western Norway. *Nordic Journal of Botany*, 34(6):645–657.

Anonymous (2018). Surfing the genomic new wave. *Nature Plants*, 4(7):393.

Arbizu, C., Ruess, H., Senalik, D., Simon, P. W., and Spooner, D. M. (2014). Phylogenomics of the carrot genus (*Daucus*, Apiaceae). *American Journal of Botany*, 101(10):1666–1685.

Atienza, S. G., Torres, A. M., Millan, T., and Cubero, J. I. (2005). Genetic diversity in *Rosa* as revealed by RAPDs. *Agriculturae Conspectus Scientificus (ACS)*, 70(3):75–85.

Atif Riaz (2011). Assessment of biodiversity based on morphological characteristics and RAPD markers among genotypes of wild rose species. *African Journal of Biotechnology*, 10(59):12520–12526.

Aubert, D. (2015). A formal analysis of phylogenetic terminology: Towards a reconsideration of the current paradigm in systematics. *Phytoneuron*, 2015-66.

Augusto Corrêa dos Santos, R., Goldman, G. H., and Riaño-Pachón, D. M. (2017). ploidyNGS: visually exploring ploidy with Next Generation Sequencing data. *Bioinformatics*, 33(16):2575–2576.

Avise, J. C. (2000). *Phylogeography: The History and Formation of Species.* Harvard University Press.

Babineau, M., Gagnon, E., and Bruneau, A. (2013). Phylogenetic utility of 19 low copy nuclear genes in closely related genera and species of caesalpinioid legumes. *South African Journal of Botany*, 89:94–105.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., and Johnson, E. A. (2008). Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLOS ONE*, 3(10):e3376.

Bakker, P., Maes, B., Maskew, R., and Stace, C. (2019). Dog-roses (*Rosa* sect. *Caninae*): towards a consensus taxonomy. *British & Irish Botany*, 1(1):7–19.

Baldwin, B. G., Goldman, D., Keil, D. J., Patterson, R., and Rosatti, T. J. (2012). *The Digital Jepson Manual: Vascular Plants of California.* University of California Press, 2nd edition.

Ballmer, D. (2018). Dogrose evolution and its implications for conservation. Master's thesis, University of Zurich, CH, Zurich, CH.

Bapteste, E., Susko, E., Leigh, J., MacLeod, D., Charlebois, R., and Doolittle, W. (2005). Do orthologous gene phylogenies really support tree-thinking? *BMC Evolutionary Biology*, 5(33).

Basaki, T., Mardi, M., Kermani, M. J., Pirseyedi, S., Ghaffari, M., Haghnazari, A., Shanjani, P. S., and Koobaz, P. (2009). Assessing *Rosa persica* genetic diversity using amplified fragment length polymorphisms analysis. *Scientia Horticulturae*, 120(4):538–543.

Bashir, T., Chandra Mishra, R., Hasan, M. M., Mohanta, T. K., and Bae, H. (2018). Effect of Hybridization on Somatic Mutations and Genomic Rearrangements in Plants. *International Journal of Molecular Sciences*, 19(12).

Baudino, S., Blerot, B., Roccia, A., and Caissard, J.-C. (2013). Les roses et la production d'huile essentielle pour la parfumerie. *Jardins de France*, 623.

Baum, D. (1992). Phylogenetic species concepts. *Trends in Ecology & Evolution*, 7(1):1–2.

Beck, J. B., Windham, M. D., Yatskievych, G., and Pryer, K. M. (2010). A Diploids-First Approach to Species Delimitation and Interpreting Polyploid Evolution in the Fern Genus *Astrolepis* (Pteridaceae). *Systematic Botany*, 35(2):223–234.

Becker, H. F. (1963). The fossil record of the genus *Rosa*. *Bulletin of the Torrey Botanical Club*, 90(2):99–110.

Bein, E., Habte, B., Jaber, A., Birnie, A., and Tengnas, B. (1996). *Useful Trees and Shrubs in Eritrea. Identification, Propagation and Management for Agricultural and Pastoral Communities.* Number 12 in Technical Handbook. Regional Soil Conservation Unit, RSCU, Nairobi, Kenya.

Benoit, L. (2019). Kenyan Roses in a Global Market. *Mambo!*, 16(8).

Bertrand, Y. J. K., Scheen, A.-C., Marcussen, T., Pfeil, B. E., de Sousa, F., and Oxelman, B. (2015). Assignment of Homoeologs to Parental Genomes in Allopolyploids for Species Tree Inference, with an Example from *Fumaria* (Papaveraceae). *Systematic Biology*, 64(3):448–471.

Bhargava, A. and Fuentes, F. F. (2010). Mutational dynamics of microsatellites. *Molecular Biotechnology*, 44(3):250–266.

Bibault, J.-E. and Tinhofer, I. (2017). The role of Next-Generation Sequencing in tumoral radiosensitivity prediction. *Clinical and Translational Radiation Oncology*, 3:16–20.

Bininda-Emonds, O. R. P. (2004). Trees versus Characters and the Supertree/Supermatrix "Paradox". *Systematic Biology*, 53(2):356–359.

Birky, C. W. (1995). Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. *Proceedings of the National Academy of Sciences*, 92(25):11331–11338.

Björklund, M. (2019). Lamarck, the Father of Evolutionary Ecology? *Trends in Ecology & Evolution*, 34(10):874–875.

Blake, W. (1794). The Sick Rose. In *Songs of Innocence and Experience*. Blake W.

Bleidorn, C. (2017). *Phylogenomics - An Introduction.* Springer International Publishing, Cham, Switzerland.

Blischak, P. D., Latvis, M., Morales-Briones, D. F., Johnson, J. C., Stilio, V. S. D., Wolfe, A. D., and Tank, D. C. (2018). Fluidigm2purc: Automated processing and haplotype inference for double-barcoded PCR amplicons. *Applications in Plant Sciences*, 6(6).

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120.

Bombarely, A. (2018). Roses for Darwin. *Nature Plants*, 4(7):406–407.

Boulenger, G. A. (1924). *Les roses d'Europe de l'Herbier Crépin: (Grande-Bretagne, France, Belgique, Pays-Bas, Suisse, Allemagne).* Goemaere.

Boulenger, G. A. (1933). Révision des Roses d'Asie de la section des *Synstylae*. *Bulletin du Jardin botanique de l'État à Bruxelles*, 9:203–279.

Boulenger, G. A. (1935). Révision des Roses d'Asie de la section des *Eglanteriae*, groupe des *Pimpinelli-Suavifoliae*, *Orientales* et *Alpinae-Vestitae*. *Bulletin du Jardin botanique de l'État à Bruxelles*, 13:165–266.

Boulenger, G. A. (1936). Révision des Roses d'Asie, section des *Eglanteriae* (suite et fin), *Chinenses*, *Bracteatae*, *Banksianae* et *Microphyllae*. *Bulletin du Jardin botanique de l'État à Bruxelles*, 14:115–221.

Bourke, P. M., Arens, P., Voorrips, R. E., Esselink, G. D., Koning-Boucoiran, C. F. S., Van't Westende, W. P. C., Santos Leonardo, T., Wissink, P., Zheng, C., van Geest, G., Visser, R. G. F., Krens, F. A., Smulders, M. J. M., and Maliepaard, C. (2017). Partial preferential chromosome pairing is genotype dependent in tetraploid rose. *The Plant Journal*, 90(2):330–343.

Bourke, P. M., Gitonga, V. W., Voorrips, R. E., Visser, R. G. F., Krens, F. A., and Maliepaard, C. (2018). Multi-environment QTL analysis of plant and flower morphological traits in tetraploid rose. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, 131(10):2055–2069.

Bouron, J. B. (2017). Le marché mondial des roses. *Géoconfluences.*

Bowler, P. J. (2003). *Evolution: The History of an Idea.* University of California Press.

Boyd, P. (2015). Scots roses and related cultivards of *Rosa spinosissima* - A review. *Acta Horticulturae*, 1064:21–30.

Bragg, J. G., Potter, S., Bi, K., and Moritz, C. (2016). Exon capture phylogenomics: efficacy across scales of divergence. *Molecular Ecology Resources*, 16(5):1059–1068.

Brandrud, M. K., Paun, O., Lorenzo, M. T., Nordal, I., and Brysting, A. K. (2017). RADseq provides evidence for parallel ecotypic divergence in the autotetraploid *Cochlearia officinalis* in Northern Norway. *Scientific Reports*, 7(5573).

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., Jovanovich, S. B., Krstic, P. S., Lindsay, S., Ling, X. S., Mastrangelo, C. H., Meller, A., Oliver, J. S., Pershin, Y. V., Ramsey, J. M., Riehn, R., Soni, G. V., Tabard-Cossa, V., Wanunu, M., Wiggin, M., and Schloss, J. A. (2008). The potential and challenges of nanopore sequencing. *Nature Biotechnology*, 26(10):1146–1153. WOS:000259926000029.

Brassac, J. and Blattner, F. R. (2015). Species-Level Phylogeny and Polyploid Relationships in *Hordeum* (Poaceae) Inferred by Next-Generation Sequencing and In Silico Cloning of Multiple Nuclear Loci. *Systematic Biology*, 64(5):792–808.

Brown, N. and Murphy, P. (2000). Neolithic and Bronze Age. In *Research and Archaeology: A framework for the Eastern Counties*, volume 2. The Scole Archaeological Committee for East Anglia, Norwich, UK.

Brumme, H., Gladis, T., Hawel, T., and Schulz, G. (2013). Die Gattung *Rosa* L. Wildrosen im Europa-Rosarium Sangerhausen.

Brummitt, R. K. (2002). How to Chop up a Tree. *Taxon*, 51(1):31–41.

Bruneau, A., Starr, J. R., and Joly, S. (2007). Phylogenetic relationships in the genus *Rosa*: new evidence from chloroplast DNA sequences and an appraisal of current knowledge. *Systematic Botany*, 32(2):366–378.

Burgess, M. B., Cushman, K. R., Doucette, E. T., Frye, C. T., and Campbell, C. S. (2015). Understanding diploid diversity: A first step in unraveling polyploid, apomictic complexity in *Amelanchier*. *American Journal of Botany*, 102(12):2041–2057.

Burnat, E. and Gremli, A. (1886). *Observations sur quelques roses de l'Italie.* H. Georg, Genève & Bâle, Lyon.

Burns, R. (1834). A Red, Red Rose. In *The Works of Robert Burns*, volume 2, page 274. The Ettrick Shepherd and William Motherwell, esq., Glasgow, UK.

Buti, M., Moretto, M., Barghini, E., Mascagni, F., Natali, L., Brilli, M., Lomsadze, A., Sonego, P., Giongo, L., Alonge, M., Velasco, R., Varotto, C., Šurbanovski, N., Borodovsky, M., Ward, J. A., Engelen, K., Cavallini, A., Cestaro, A., and Sargent, D. J. (2018). The genome sequence and transcriptome of *Potentilla micrantha* and their comparison to *Fragaria vesca* (the woodland strawberry). *GigaScience*, 7(4).

Cabrera, A., Kozik, A., Howad, W., Arus, P., Iezzoni, A. F., and Knaap, E. (2009). Development and bin mapping of a Rosaceae Conserved Ortholog Set (COS) of markers. *BMC Genomics*, 10(562).

Cai, D., Rodríguez, F., Teng, Y., Ané, C., Bonierbale, M., Mueller, L. A., and Spooner, D. M. (2012). Single copy nuclear gene analysis of polyploidy in wild potatoes (*Solanum* section *Petota*). *BMC Evolutionary Biology*, 12(70):70.

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

Canback, B., Andersson, S. G. E., and Kurland, C. G. (2002). The global phylogeny of glycolytic enzymes. *Proceedings of the National Academy of Sciences*, 99(9):6097–6102.

Caser, M. (2017). Pollen grains and tubes. In *Reference Module in Life Sciences*. Elsevier.

Casillas, S. and Barbadilla, A. (2017). Molecular Population Genetics. *Genetics*, 205(3):1003–1035.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, 17(4):540–552.

Cavalier-Smith, T. (1982). The origins of plastids. *Biological Journal of the Linnean Society*, 17(3):289–306.

Charlesworth, B. and Charlesworth, D. (2017). Population genetics from 1966 to 2016. *Heredity*, 118(1):2–9.

Chen, X., Liu, Y., Sun, J., Wang, L., and Zhou, S. (2019). The complete chloroplast genome sequence of *Rosa acicularis* in Rosaceae. *Mitochondrial DNA Part B*, 4(1):1743–1744.

Cheng, S., Melkonian, M., Smith, S. A., Brockington, S., Archibald, J. M., Delaux, P.-M., Li, F.-W., Melkonian, B., Mavrodiev, E. V., Sun, W., Fu, Y., Yang, H., Soltis, D. E., Graham, S. W., Soltis, P. S., Liu, X., Xu, X., and Wong, G. K.-S. (2018). 10kp: A phylodiverse genome sequencing plan. *GigaScience*, 7(3).

Chery, J. G., Sass, C., and Specht, C. D. (2017). Development of single-copy nuclear intron markers for species-level phylogenetics: Case study with Paullinieae (Sapindaceae). *Applications in Plant Sciences*, 5(9).

Chinnappareddy, L. R. D., Khandagale, K., Srinivas Reddy, S. H., Kanupriya, C., Chennareddy, A., and Singh, T. H. (2012). SSR-Based DNA Barcodes as a Tool for Identification of Eggplant Genotypes. *International Journal of Vegetable Science*, 18(3):260–271.

Choi, B., Crisp, M. D., Cook, L. G., Meusemann, K., Edwards, R. D., Toon, A., and Külheim, C. (2019). Identifying genetic markers for a range of phylogenetic utility–From species to family level. *PLoS ONE*, 14(8).

Christ, H. (1873). *Die Rosen der Schweiz mit Berücksichtigung der umliegenden Gebiete Mittel- und Süd-Europa's: Ein monographischer Versuch.* Georg.

Chwalkowski, F. (2016). Rose: the love of freedom. In *Symbols in Arts, Religion and Culture: The Soul of Nature*, pages 211–220. Cambridge Scholars Publishing, Newcastle upon Tyne, UK.

Clark, L. V. and Jasieniuk, M. (2011). polysat: an R package for polyploid microsatellite analysis. *Molecular Ecology Resources*, 11(3):562–566.

Cole, P. and Melton, B. (1986). Self- and cross-compatibility relationships among genotypes and between ploidy of the rose. *Journal of the American Society for Horticultural Science*, 111:122–125.

Conington, J. (1872). XV. *Jam pauca aratro.* In *Horace. The Odes and Carmen Saeculare of Horace*, page 55. George Bell and Sons, London, UK, 5th edition.

Cornides, E. (1967). Rose und Schwert im päpstlichen Zeremoniell. In *Wiener Dissertationen aus dem Gebiete der Geschichte*, volume 9, page 182. Wissenschaftliches Antiquariat H. Geyer.

Corriveau, J. L. and Coleman, A. W. (1988). Rapid screening method to detect potential biparental inheritance of plastid DNA and results for over 200 angiosperm species. *American Journal of Botany*, 75(10):1443–1458.

Crhak Khaitova, L., Werlemark, G., Kovarikova, A., Nybom, H., and Kovarik, A. (2014). High Penetrance of a Pan-*Canina* Type rDNA Family in Intersection *Rosa* Hybrids Suggests Strong Selection of Bivalent Chromosomes in the Section *Caninae*. *Cytogenetic and Genome Research*, 143(1-3):104–113.

Crumpton-Taylor, M., Grandison, S., Png, K. M. Y., Bushby, A. J., and Smith, A. M. (2012). Control of starch granule numbers in *Arabidopsis* chloroplasts. *Plant Physiology*, 158(2):905–916.

Crépin, F. (1889). Sketch of a new Classification of Roses. *Journal of the Royal Horticultural Society*, 11(3):217–228.

Crépin, F. (1891). Nouvelle classification des Roses. *Journal des Roses*, 15:41–43, 53–55, 76–77.

Cuizhi, G. and Robertson, K. R. (2003). 41. ROSA Linnaeus, Sp. Pl. 1: 491. 1753. In *Flora of China. Vol. 9 (Pittosporaceae through Connaraceae)*, volume 9, pages 339–381. Science Press, Beijing, and Missouri Botanical Garden Press, St. Louis.

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and high-performance computing. *Nature methods*, 9(8):772.

Darwin, C. (1866). *On the Origin of Species.* J. Murray, London, UK, 4 edition.

De Cock, K., Vander Mijnsbrugge, K., Breyne, P., Van Bockstaele, E., and Van Slycken, J. (2008). Morphological and AFLP-based differentiation within the taxonomical complex section *Caninae* (subgenus *Rosa*). *Annals of Botany*, 102(5):685–697.

de la Roche, G., Rowley, G. D., Lawalrée, A., and Stearn, W. T. (1976). *Commentaries to "Les Roses" by P. J. Redouté: a contribution to the history of the genus Rosa / Commentaires sur "Les Roses" de P. J. Redouté : Apport à l'histoire du genre Rosa.* De Schutter, Antwerpen.

De Queiroz, K. (2007). Species Concepts and Species Delimitation. *Systematic Biology*, 56(6):879–886.

de Queiroz, K. and Good, D. A. (1997). Phenetic Clustering in Biology: A Critique. *The Quarterly Review of Biology*, 72(1):3–30.

De Riek, J., De Cock, K., Smulders, M. J., and Nybom, H. (2013). AFLP-based population structure analysis as a means to validate the complex taxonomy of dogroses (*Rosa* section *Caninae*). *Molecular Phylogenetics and Evolution*, 67(3):547–559.

de Saint-Exupéry, A. (1943). *Le Petit Prince*. Reynal and Hitchcock, New York, NY, USA, 1 edition.

de Wet, J. M. J. (1971). Polyploidy and Evolution in Plants. *Taxon*, 20(1):29–35.

Debener, T., Bretzke, M., Dreier, K., Spiller, M., Linde, M., Kaufmann, H., Berger, R., and Krings, U. (2010). Genetic and molecular analyses of key loci involved in self incompatibility and floral scent in roses. *Acta Horticulturae*, 870:183–190.

Debray, K., Marie-Magdelaine, J., Ruttink, T., Clotault, J., Foucher, F., and 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 Evolutionary Biology*, 19(152).

Degnan, J. H. and Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, 24(6):332–340.

Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, 6(5):361–375.

Deng, H., Zhang, G.-Q., Lin, M., Wang, Y., and Liu, Z.-J. (2015). Mining from transcriptomes: 315 single-copy orthologous genes concatenated for the phylogenetic analyses of Orchidaceae. *Ecology and Evolution*, 5(17):3800–3807.

Deplazes-Zemp, A., Abiven, S., Schaber, P., Schaepman, M., Schaepman-Strub, G., Schmid, B., Shimizu, K. K., and Altermatt, F. (2018). The Nagoya Protocol could backfire on the Global South. *Nature Ecology & Evolution*, 2(6):917–919.

DeVore, M. L. and Pigg, K. B. (2007). A brief review of the fossil history of the family Rosaceae with a focus on the Eocene Okanogan Highlands of eastern Washington State, USA, and British Columbia, Canada. *Plant Systematics and Evolution*, 266(1-2):45–57.

Dierckxsens, N., Mardulyn, P., and Smits, G. (2017). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research*, 45(4):e18.

Dillenberger, M. S., Wei, N., Tennessen, J. A., Ashman, T.-L., and Liston, A. (2018). Plastid genomes reveal recurrent formation of allopolyploid *Fragaria*. *American Journal of Botany*, 105(5):862–874.

DiMeglio, L. M., Staudt, G., Yu, H., and Davis, T. M. (2014). A Phylogenetic Analysis of the Genus *Fragaria* (Strawberry) Using Intron-Containing Sequence from the ADH-1 Gene. *PLoS ONE*, 9(7).

Dobson, H. E. M., Danielson, E. M., and Wesep, I. D. V. (1999). Pollen odor chemicals as modulators of bumble bee foraging on *Rosa rugosa* Thunb. (Rosaceae). *Plant Species Biology*, 14(2):153–166.

Dong, W., Xu, C., Cheng, T., and Zhou, S. (2013). Complete Chloroplast Genome of *Sedum sarmentosum* and Chloroplast Genome Evolution in Saxifragales. *PLoS ONE*, 8(10):e77965.

Donoghue, M. J. (1985). A Critique of the Biological Species Concept and Recommendations for a Phylogenetic Alternative. *The Bryologist*, 88(3):172–181.

Douglas, S. E. (1998). Plastid evolution: origins, diversity, trends. *Current Opinion in Genetics & Development*, 8(6):655–661.

Doyle, J. J. (1992). Gene Trees and Species Trees: Molecular Systematics as One-Character Taxonomy. *Systematic Botany*, 17(1):144–163.

Duarte, J. M., Wall, P. K., Edger, P. P., Landherr, L. L., Ma, H., Pires, P. K., Leebens-Mack, J., and others (2010). Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evolutionary Biology*, 10(61).

Dubois, A. (2011). Species and “strange species” in zoology: Do we need a “unified concept of species”? *Comptes Rendus Palevol*, 10(2-3):77–94.

Dumortier, B. C. (1824). *Notice sur un nouveau genre de plantes : Hulthemia; précédée d'un aperçu sur la classification des roses.* Imprimerie de J. Casterman, Aîné, Tournay, Belgium.

Díaz-Pérez, A., López-Álvarez, D., Sancho, R., and Catalán, P. (2018). Reconstructing the origins and the biogeography of species' genomes in the highly reticulate allopolyploid-rich model grass genus *Brachypodium* using minimum evolution, coalescence and maximum likelihood approaches. *Molecular Phylogenetics and Evolution*, 127:256–271.

Earl, D. A. and vonHoldt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, 4(2):359–361.

Eaton, D. A. R. and Ree, R. H. (2013). Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: Orobanchaceae). *Systematic Biology*, 62(5):689–706.

Edelman, D. (1975). The Eocene germer basin flora of South-Central Idaho. Master's thesis, University of Idaho, Moscow, Idaho, USA.

Edger, P. P., VanBuren, R., Colle, M., Poorten, T. J., Wai, C. M., Niederhuth, C. E., Alger, E. I., Ou, S., Acharya, C. B., Wang, J., Callow, P., McKain, M. R., Shi, J., Collier, C., Xiong, Z., Mower, J. P., Slovin, J. P., Hytönen, T., Jiang, N., Childs, K. L., and Knapp, S. J. (2018). Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (*Fragaria vesca*) with chromosome-scale contiguity. *GigaScience*, 7(2).

Eickbush, T. H. and Eickbush, D. G. (2007). Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics*, 175(2):477–485.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, 323(5910):133–138.

Eisen, J. A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research*, 8:163–167.

Eisen, J. A. and Hanawalt, P. C. (1999). A phylogenomic study of DNA repair genes, proteins, and processes. *Mutation Research/DNA Repair*, 435(3):171–213.

Eisen, J. A., Kaiser, D., and Myers, R. M. (1997). Gastrogenomic delights: a movable feast. *Nature Medicine*, 3(10):1076–1078.

Elburg, R. (2010). A Neolithic treasure chest. *The European Archaeologist*, 33:4–6.

Enright, A. J., Dongen, S. V., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575–1584.

Ertter, B. and Lewis, W. H. (2016). Relationships, infrataxa, and hybrids of *Rosa gymnocarpa* (Rosaceae). *Madroño*, 63(3):268–280.

Escudero, M., Eaton, D. A. R., Hahn, M., and Hipp, A. L. (2014). Genotyping-by-sequencing as a tool to infer phylogeny and ancestral hybridization: a case study in *Carex* (Cyperaceae). *Molecular Phylogenetics and Evolution*, 79:359–367.

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, 14(8):2611–2620.

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., and Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, 61(5):717–726.

Farris, J. S., Källersjö, M., Kluge, A. G., and Bult, C. (1995). Testing significance of incongruence. *Cladistics*, 10(3):315–319.

Favre, A., Päckert, M., Pauls, S. U., Jähnig, S. C., Uhl, D., Michalak, I., and Muellner-Riehl, A. N. (2015). The role of the uplift of the Qinghai-Tibetan Plateau for the evolution of Tibetan biotas. *Biological Reviews*, 90(1):236–253.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution; International Journal of Organic Evolution*, 39(4):783–791.

Ferrand, N. (2015). *Créateurs de roses. A la conquête des marchés (1820-1939).* La Pierre et l'Ecrit. PUG, 1st edition.

Fitch, W. M. (1970). Distinguishing Homologous from Analogous Proteins. *Systematic Zoology*, 19(2):99–113.

Fougère-Danezan, M., Joly, S., Bruneau, A., Gao, X.-F., and Zhang, L.-B. (2015). Phylogeny and biogeography of wild roses with specific attention to polyploids. *Annals of Botany*, 115(2):275–291.

Foxton-Smythe, R. (2013). The History and Cultivation of Roses. http://www.houseplantsguru.com/the-history-and-cultivation-of-roses, accessed December 12, 2019.

Gallardo, M. H. (2017). Phylogenetics, Reticulation and Evolution. In Abdurakhmonov, I. Y., editor, *Phylogenetics*. InTech.

Galtier, N. and Daubin, V. (2008). Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1512):4023–4029.

Gao, Y.-D., Gao, X.-F., and Harris, A. (2019). Species Boundaries and Parapatric Speciation in the Complex of Alpine Shrubs, *Rosa sericea* (Rosaceae), Based on Population Genetics and Ecological Tolerances. *Frontiers in Plant Science*, 10(321).

Gao, Y.-D., Zhang, Y., Gao, X.-F., and Zhu, Z.-M. (2015). Pleistocene glaciations, demographic expansion and subsequent isolation promoted morphological heterogeneity: A phylogeographic study of the alpine *Rosa sericea* complex (Rosaceae). *Scientific Reports*, 5(11698).

Gar, O., Sargent, D. J., Tsai, C.-J., Pleban, T., Shalev, G., Byrne, D. H., and Zamir, D. (2011). An Autotetraploid Linkage Map of Rose (*Rosa hybrida*) Validated Using the Strawberry (*Fragaria vesca*) Genome Sequence. *PLoS ONE*, 6(5).

García-Pereira, M. J., Caballero, A., and Quesada, H. (2010). Evaluating the Relationship between Evolutionary Divergence and Phylogenetic Accuracy in AFLP Data Sets. *Molecular Biology and Evolution*, 27(5):988–1000.

Gatesy, J., Baker, R. H., and Hayashi, C. (2004). Inconsistencies in Arguments for the Supertree Approach: Supermatrices versus Supertrees of *Crocodylia*. *Systematic Biology*, 53(2):342–355.

GBIF (2019). *Rosa mollis* Sm. https://www.gbif.org/species/3003048, accessed December 12, 2019.

Gitzendanner, M. A., Soltis, P. S., Wong, G. K.-S., Ruhfel, B. R., and Soltis, D. E. (2018). Plastid phylogenomic analysis of green plants: a billion years of evolutionary history. *American Journal of Botany*, 105(3):291–301.

Glover, N. M., Redestig, H., and Dessimoz, C. (2016). Homoeologs: What Are They and How Do We Infer Them? *Trends in Plant Science*, 21(7):609–621.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351.

Gori, K., Suchan, T., Alvarez, N., Goldman, N., and Dessimoz, C. (2016). Clustering genes of common evolutionary history. *Molecular Biology and Evolution*, 33(6):1590–1605.

Granados Mendoza, C., Naumann, J., Samain, M.-S., Goetghebeur, P., De Smet, Y., and Wanke, S. (2015). A genome-scale mining strategy for recovering novel rapidly-evolving nuclear single-copy genes for addressing shallow-scale phylogenetics in *Hydrangea*. *BMC Evolutionary Biology*, 15(132).

Grimm, J. and Grimm, W. (1884). Little Briar Rose. In *Household Tales*, volume 1, pages 197–200. George Bell and Sons, London, UK, Margaret Hunt, translator edition.

Grünwald, N. J., Everhart, S. E., Knaus, B. J., and Kamvar, Z. N. (2017). Best Practices for Population Genetic Analyses. *Phytopathology*, 107(9):1000–1010.

Guindon, S., Gascuel, O., and Rannala, B. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5):696–704.

Guoliang, W. (2003). History of roses in cultivation / Ancient Chinese roses. In *Encyclopedia of Rose science*, pages 387–395. A.V. Roberts, T. Debener and S. Gudin (Eds.), Oxford, Elsevier edition.

Hably, L., Kvacek, Z., and Manchester, S. (2000). Shared taxa of land plants in the Oligocene of Europe and North America in context of Holarctic phytogeography. *Acta Universitatis Carolinae Geologica*, 44:59–74.

Hamrick, J. L. and Godt, M. J. W. (1990). Allozyme diversity in plant species. *Plant population genetics, breeding, and genetic resources.*, pages 43–63.

Han, F., Peng, Y., Xu, L., and Xiao, P. (2014). Identification, characterization, and utilization of single copy genes in 29 angiosperm genomes. *BMC Genomics*, 15(504).

Hanks, G. (2018). A review of production statistics for the cut flower and foliage sector. Updated 2018. (part of AHDB Horticulture project PO BOF 002a). Technical report, The National Cut Flower Centre, AHDB.

Hausdorf, B. (2011). Progress Toward a General Species Concept. *Evolution*, 65(4):923–931.

Heitzler, P. (2015). *Rosa persica* and Its Descendants. *Newsletter of the World Federation of Rose Societies*, 12:14–18.

Henker, D. H. (2000). *Rosa*. In *Illustrierte Flora von Mitteleuropa*, volume 4(2C), pages 1–108. Parey, Berlin, Germany, 2nd edition.

Hennig, W. (1966). *Phylogenetic Systematics*. University of Illinois Press.

Herklotz, V. and Ritz, C. M. (2017). Multiple and asymmetrical origin of polyploid dog rose hybrids (*Rosa* L. sect. *Caninae* (DC.) Ser.) involving unreduced gametes. *Annals of Botany*, 120(2):209–220.

Herrera, S. and Shank, T. M. (2016). RAD sequencing enables unprecedented phylogenetic resolution and objective species delimitation in recalcitrant divergent taxa. *Molecular Phylogenetics and Evolution*, 100:70–79.

Hibrand Saint-Oyant, L., Crespel, L., Rajapakse, S., Zhang, L., and Foucher, F. (2008). Genetic linkage maps of rose constructed with new microsatellite markers and locating QTL controlling flowering traits. *Tree Genetics & Genomes*, 4(1):11.

Hibrand Saint-Oyant, L., Ruttink, T., Hamama, L., Kirov, I., Lakhwani, D., Zhou, N. N., Bourke, P. M., Daccord, N., Leus, L., Schulz, D., Geest, H. V. d., Hesselink, T., Laere, K. V., Debray, K., Balzergue, S., Thouroude, T., Chastellier, A., Jeauffre, J., Voisine, L., Gaillard, S., Borm, T. J. A., Arens, P., Voorrips, R. E., Maliepaard, C., Neu, E., Linde, M., Paslier, M. C. L., Bérard, A., Bounon, R., Clotault, J., Choisne, N., Quesneville, H., Kawamura, K., Aubourg, S., Sakr, S., Smulders, M. J. M., Schijlen, E., Bucher, E., Debener, T., Riek, J. D., and Foucher, F. (2018). A high-quality genome sequence of *Rosa chinensis* to elucidate ornamental traits. *Nature Plants*, 4(7):473–484.

Hillis, D. M. and Bull, J. J. (1993). An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis. *Systematic Biology*, 42(2):182–192.

Hilu, K. W., Black, C. M., and Oza, D. (2014). Impact of gene molecular evolution on phylogenetic reconstruction: a case study in the Rosids (Superorder Rosanae, Angiosperms). *PLoS ONE*, 9(6).

Hipp, A. L., Eaton, D. A. R., Cavender-Bares, J., Fitzek, E., Nipper, R., and Manos, P. S. (2014). A framework phylogeny of the American oak clade based on sequenced RAD data. *PloS One*, 9(4):e93975.

Hirakawa, H., Shirasawa, K., Kosugi, S., Tashiro, K., Nakayama, S., Yamada, M., Kohara, M., Watanabe, A., Kishida, Y., Fujishiro, T., Tsuruoka, H., Minami, C., Sasamoto, S., Kato, M., Nanri, K., Komaki, A., Yanagi, T., Guoxin, Q., Maeda, F., Ishikawa, M., Kuhara, S., Sato, S., Tabata, S., and Isobe, S. N. (2014). Dissection of the octoploid strawberry genome by deep sequencing of the genomes of *Fragaria* species. *DNA Research*, 21(2):169–181.

Hirota, S. K., Miki, N., Yasumoto, A. A., and Yahara, T. (2018). UV bullseye contrast of *Hemerocallis* flowers attracts hawkmoths but not swallowtail butterflies. *Ecology and Evolution*, 9(1):52–64.

Hoquet, T. (2007). *Buffon/Linné - Éternels rivaux de la biologie ?* Quai des Sciences. Dunod, Paris, France.

Hort, A. (1916). *Enquiry into plants and minor works on odours and weathers, with an English translation by Sir Arthur Hort, Bart., M.A.* W. Heinemann.

Huber, K. T. and Moulton, V. (2006). Phylogenetic networks from multi-labelled trees. *Journal of Mathematical Biology*, 52(5):613–632.

Huelsenbeck, J. and Bull, J. J. (1996). A likelihood ratio test to detect conflicting phylogenetic signal. *Systematic Biology*, 45(1):92–98.

Huelsenbeck, J. P. and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755.

Hughes, C. E., Eastwood, R. J., and Bailey, C. D. (2006). From famine to feast? Selecting nuclear DNA sequence loci for plant species-level phylogeny reconstruction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1465):211–225.

Hurst, C. C. (1925). Chromosomes and characters in *Rosa* and their significance in the origin of species. *Experiments in Genetics*, 37:534–550.

Hurst, C. C. (1928). Differential polyploidy in the genus *Rosa* L. *Zeitschrift fur induktive Abstammungs und Vererbungslehre*, Supplement 2:868–906.

Hurst, C. C. (1941). Notes on the origin and evolution of our garden roses. *Journal of the Royal Horticultural Society*, 66:73–82, 242–250, 282–289.

Huson, D. H. and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2):254–267.

Huson, D. H. and Scornavacca, C. (2012). Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks. *Systematic Biology*, 61(6):1061–1067.

Huxley, J. (1943). *Evolution: The Modern Synthesis.* Harper & brothers, New York, NY, USA and London, UK.

Häuser, C. L. (1987). The debate about the biological species concept - a review. *Journal of Zoological Systematics and Evolutionary Research*, 25(4):241–257.

Hörandl, E. (2006). Paraphyletic versus monophyletic taxa—evolutionary versus cladistic classifications. *Taxon*, 55(3):564–570.

Igea, J., Juste, J., and Castresana, J. (2010). Novel intron markers to study the phylogeny of closely related mammalian species. *BMC Evolutionary Biology*, 10(369).

Iguchi, A., Yoshioka, Y., Forsman, Z. H., Knapp, I. S., Toonen, R. J., Hongo, Y., Nagai, S., and Yasuda, N. (2019). RADseq population genomics confirms divergence across closely related species in blue coral (*Heliopora coerulea*). *BMC Evolutionary Biology*, 19(187).

Iwata, H., Kato, T., and Ohno, S. (2000). Triparental origin of Damask roses. *Gene*, 259(1):53–59.

Jackman, S. D., Vandervalk, B. P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S. A., Jahesh, G., Khan, H., Coombe, L., Warren, R. L., and Birol, I. (2017). ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome Research*, 27(5):768–777.

Jacques, F., Su, T., Spicer, R. A., Xing, Y.-W., Huang, Y.-J., and Zhou, Z.-K. (2014). Late Miocene southwestern Chinese floristic diversity shaped by the southeastern uplift of the Tibetan Plateau. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 411:208–215.

Jan, C. H., Byrne, D. H., Manhart, J., and Wilson, H. (1999). Rose germplasm analysis with RAPD markers. *HortScience*, 34(2):341–345.

Jarvis, C. E. (1992). Seventy-two proposals for the conservation of types of selected Linnaean generic names, the report of subcommittee 3c on the lectotypification of Linnaean generic names. *Taxon*, 41:568–569.

Jeffroy, O., Brinkmann, H., Delsuc, F., and Philippe, H. (2006). Phylogenomics: the beginning of incongruence? *Trends in Genetics*, 22(4):225–231.

Jeon, J.-H. and Kim, S.-C. (2019). Comparative analysis of the complete chloroplast genome sequences of three closely related East-Asian wild roses (*Rosa* sect. *Synstylae*; Rosaceae). *Genes*, 10(1):23.

Jian, H., Zhang, H., Tang, K., Li, S., Wang, Q., Zhang, T., Qiu, X., and Yan, H. (2010). Decaploidy in *Rosa praelucens* Byhouwer (Rosaceae) Endemic to Zhongdian Plateau, Yunnan, China. *Caryologia*, 63(2):162–167.

Jian, H., Zhang, T., Wang, Q., Yan, H., Qiu, X., Zhou, N., Li, S., Chen, M., Zhang, H., and Tang, K. (2014). Nuclear DNA content and 1cx-value variations in genus *Rosa* L. *Caryologia*, 67(4):273–280.

Jian, H.-Y., Zhang, Y.-H., Yan, H.-J., Qiu, X.-Q., Wang, Q.-G., Li, S.-B., and Zhang, S.-D. (2018). The complete chloroplast genome of a key ancestor of modern roses, *Rosa chinensis* var. *spontanea*, and a comparison with congeneric species. *Molecules (Basel, Switzerland)*, 23(2).

Jiao, Y. and Guo, H. (2014). Prehistory of the Angiosperms. In *Advances in Botanical Research*, volume 69, pages 223–245. Elsevier.

Jin, G., Nakhleh, L., Snir, S., and Tuller, T. (2006). Maximum likelihood of phylogenetic networks. *Bioinformatics*, 22(21):2604–2611.

Jin, J.-J., Yu, W.-B., Yang, J.-B., Song, Y., Yi, T.-S., and Li, D.-Z. (2018). GetOrganelle: a simple and fast pipeline for de novo assembly of a complete circular chloroplast genome using genome skimming data. *bioRχiv*, 256479.

Johnson, M. T. J., Carpenter, E. J., Tian, Z., Bruskiewich, R., Burris, J. N., Carrigan, C. T., Chase, M. W., Clarke, N. D., Covshoff, S., dePamphilis, C. W., Edger, P. P., Goh, F., Graham, S., Greiner, S., Hibberd, J. M., Jordon-Thaden, I., Kutchan, T. M., Leebens-Mack, J., Melkonian, M., Miles, N., Myburg, H., Patterson, J., Pires, J. C., Ralph, P., Rolf, M., Sage, R. F., Soltis, D., Soltis, P., Stevenson, D., Jr, C. N. S., Surek, B., Thomsen, C. J. M., Villarreal, J. C., Wu, X., Zhang, Y., Deyholos, M. K., and Wong, G. K.-S. (2012). Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes. *PLoS ONE*, 7(11).

Joly, S. and Bruneau, A. (2007). Delimiting species boundaries in *Rosa* sect. *Cinnamomeae* (Rosaceae) in Eastern North America. *Systematic Botany*, 32(4):819–836.

Joly, S., Bruneau, A., and Baker, A. (2006a). Incorporating allelic variation for reconstructing the evolutionary history of organisms from multiple genes: An example from *Rosa* in North America. *Systematic Biology*, 55(4):623–636.

Joly, S., Starr, J. R., Lewis, W. H., and Bruneau, A. (2006b). Polyploid and hybrid evolution in roses east of the Rocky Mountains. *American Journal of Botany*, 93(3):412–425.

Jordan, C. Y., Lohse, K., Turner, F., Thomson, M., Gharbi, K., and Ennos, R. A. (2018). Maintaining their genetic distance: little evidence for introgression between widely hybridizing species of *Geum* with contrasting mating systems. *Molecular Ecology*, 27(5):1214–1228.

Joret, C. (1892). *La Rose dans l'Antiquité et au Moyen Age. Histoire, légendes et symbolisme.* Émile Bouillon, Paris, France.

Joyaux, F. (2015). *Nouvelle encyclopédie des roses anciennes*. Eugen Ulmer, Paris, France.

Jukes, T. H. and Cantor, C. R. (1969). CHAPTER 24 - Evolution of Protein Molecules. In Munro, H. N., editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press.

Kamneva, O. K., Syring, J., Liston, A., and Rosenberg, N. A. (2017). Evaluating allopolyploid origins in strawberries (*Fragaria*) using haplotypes generated from target capture sequencing. *BMC Evolutionary Biology*, 17(180).

Kates, H. R., Soltis, P. S., and Soltis, D. E. (2017). Evolutionary and domestication history of *Cucurbita* (pumpkin and squash) species inferred from 44 nuclear loci. *Molecular Phylogenetics and Evolution*, 111:98–109.

Katoh, K. and Standley, D. M. (2013). MAFFT Multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780.

Katz, M. E., Miller, K. G., Wright, J. D., Wade, B. S., Browning, J. V., Cramer, B. S., and Rosenthal, Y. (2008). Stepwise transition from the Eocene greenhouse to the Oligocene icehouse. *Nature Geoscience*, 1(5):329–334.

Kawamura, K., Hibrand-Saint Oyant, L., Crespel, L., Thouroude, T., Lalanne, D., and Foucher, F. (2011). Quantitative trait loci for flowering time and inflorescence architecture in rose. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, 122(4):661–675.

Keb-Llanes, M., González, G., Chi-Manzanero, B., and Infante, D. (2002). A rapid and simple method for small-scale DNA extraction in *Agavaceae* and other tropical plants. *Plant Molecular Biology Reporter*, 20(3):299–299.

Kellner, A., Benner, M., Walther, H., Kunzmann, L., Wissemann, V., and Ritz, C. M. (2012a). Leaf architecture of extant species of *Rosa* L. and the paleogene species *Rosa lignitum* Heer (Rosaceae). *International Journal of Plant Sciences*, 173(3):239–250.

Kellner, A., Ritz, C. M., and Wissemann, V. (2012b). Hybridization with invasive *Rosa rugosa* threatens the genetic integrity of native *Rosa mollis*. *Botanical Journal of the Linnean Society*, 170(3):472–484.

Kellner, A., Ritz, C. M., and Wissemann, V. (2014). Low genetic and morphological differentiation in the European species complex of *Rosa sherardii*, *R. mollis* and *R. villosa* (*Rosa* section *Caninae* subsection *Vestitae*). *Botanical Journal of the Linnean Society*, 174(2):240–256.

Kevan, P. G., Eisikowitch, D., Ambrose, J. D., and Kemp, J. R. (1990). Cryptic dioecy and insect pollination in *Rosa setigera* Michx. (Rosaceae), a rare plant of Carolinian Canada. *Biological Journal of the Linnean Society*, 40(3):229–243.

Kim, Y., Heo, K.-I., Nam, S., Xi, H., Lee, S., and Park, J. (2019). The complete chloroplast genome of candidate new species from *Rosa rugosa* in Korea (Rosaceae). *Mitochondrial DNA Part B*, 4(2):2433–2435.

Kinosian, S. P., Testo, W. L., Chambers, S. M., and Sessa, E. B. (2019). Using RAD Data to Confirm Parentage of Polyploids in a Reticulate Complex of Ferns. *American Fern Journal*, 109(3):267–282.

Klastersky, I. (1968). 10. *Rosa* L. In *Flora Europaea: Rosaceae to Umbelliferae*, volume 2, pages 25–33. Tutin T. G. et al. eds, Cambridge University Press edition.

Klopfstein, S., Kropf, C., and Quicke, D. L. J. (2010). An Evaluation of phylogenetic informativeness profiles and the molecular phylogeny of Diplazontinae (Hymenoptera, Ichneumonidae). *Systematic Biology*, 59(2):226–241.

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., and Mardis, E. (2013). The Next-Generation Sequencing Revolution and Its Impact on Genomics. *Cell*, 155(1):27–38.

Koopman, W. J., Wissemann, V., De Cock, K., Van Huylenbroeck, J., De Riek, J., Sabatino, G. J., Visser, D., Vosman, B., Ritz, C. M., Maes, B., and others (2008). AFLP markers as a tool to reconstruct complex relationships: a case study in *Rosa* (Rosaceae). *American Journal of Botany*, 95(3):353–366.

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., and Mayrose, I. (2015). Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources*, 15(5):1179–1191.

Kovacheva, N., Rusanov, K., and Atanassov, I. (2010). Industrial Cultivation of Oil Bearing Rose and Rose Oil Production in Bulgaria During 21st Century, Directions and Challenges. *Biotechnology & Biotechnological Equipment*, 24(2):1793–1798.

Kozik, A., Chan, B., and Michelmore, R. (2005). Tcl/Tk NCBI BLAST PARSER. University of California, Davis, https://cgpdb.ucdavis.edu/BlastParser/Blast_Parser.html, accessed December 12, 2019.

Krehenwinkel, H., Wolf, M., Lim, J. Y., Rominger, A. J., Simison, W. B., and Gillespie, R. G. (2017). Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Scientific Reports*, 7(17668).

Kruglyak, S., Durrett, R., Schug, M. D., and Aquadro, C. F. (2000). Distribution and abundance of microsatellites in the yeast genome can be explained by a balance between slippage events and point mutations. *Molecular Biology and Evolution*, 17(8):1210–1219.

Kubatko, L. S., Carstens, B. C., and Knowles, L. L. (2009). STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, 25(7):971–973.

Kuhner, M. K. and Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, 11(3):459–468.

Kumar, K. R., Cowley, M. J., and Davis, R. L. (2019). Next-Generation Sequencing and Emerging Technologies. *Seminars in Thrombosis and Hemostasis*.

Kumar, S., Filipski, A. J., Battistuzzi, F. U., Kosakovsky Pond, S. L., and Tamura, K. (2012). Statistics and Truth in Phylogenomics. *Molecular Biology and Evolution*, 29(2):457–472.

Lampridius, A. (400 AD). *Vita Antonini Heliogabali*, XXI.5. In *Historia Augusta*. na.

Lanfear, R., Calcott, B., Kainer, D., Mayer, C., and Stamatakis, A. (2014). Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evolutionary Biology*, 14(82).

Larson, E. J. (2006). *Evolution: The Remarkable History of a Scientific Theory*. Random House Publishing Group.

Latta, R. G., Bekele, W. A., Wight, C. P., and Tinker, N. A. (2019). Comparative linkage mapping of diploid, tetraploid, and hexaploid *Avena* species suggests extensive chromosome rearrangement in ancestral diploids. *Scientific Reports*, 9(1):1–12.

Le Comber, S. C., Ainouche, M. L., Kovarik, A., and Leitch, A. R. (2010). Making a functional diploid: from polysomic to disomic inheritance. *The New Phytologist*, 186(1):113–122.

Le Rougetel, H. (1988). (5) The Rose of England. *RSA Journal*, 136(5386):742–744.

Leaché, A. D., Banbury, B. L., Felsenstein, J., de Oca, A. n.-M., and Stamatakis, A. (2015). Short Tree, Long Tree, Right Tree, Wrong Tree: New Acquisition Bias Corrections for Inferring SNP Phylogenies. *Systematic Biology*, 64(6):1032–1047.

Lebel, M., Obolski, U., Hadany, L., and Sapir, Y. (2017). Pollinator-mediated selection on floral size and tube color in Linum pubescens: Can differential behavior and preference in different times of the day maintain dimorphism? *Ecology and Evolution*, 8(2):1096–1106.

Lee, A. (1913). Our imitation bismarcks. *Metropolitan Magazine*, 38(1):63.

Leebens-Mack, J. H., Barker, M. S., Carpenter, E. J., Deyholos, M. K., Gitzendanner, M. A., Graham, S. W., Grosse, I., Li, Z., Melkonian, M., Mirarab, S., Porsch, M., Quint, M., Rensing, S. A., Soltis, D. E., Soltis, P. S., Stevenson, D. W., Ullrich, K. K., Wickett, N. J., DeGironimo, L., Edger, P. P., Jordon-Thaden, I. E., Joya, S., Liu, T., Melkonian, B., Miles, N. W., Pokorny, L., Quigley, C., Thomas, P., Villarreal, J. C., Augustin, M. M., Barrett, M. D., Baucom, R. S., Beerling, D. J., Benstein, R. M., Biffin, E., Brockington, S. F., Burge, D. O., Burris, J. N., Burris, K. P., Burtet-Sarramegna, V., Caicedo, A. L., Cannon, S. B., Çebi, Z., Chang, Y., Chater, C., Cheeseman, J. M., Chen, T., Clarke, N. D., Clayton, H., Covshoff, S., Crandall-Stotler, B. J., Cross, H., dePamphilis, C. W., Der, J. P., Determann, R., Dickson, R. C., Di Stilio, V. S., Ellis, S., Fast, E., Feja, N., Field, K. J., Filatov, D. A., Finnegan, P. M., Floyd, S. K., Fogliani, B., García, N., Gâteblé, G., Godden, G. T., Goh, F. Q. Y., Greiner, S., Harkess, A., Heaney, J. M., Helliwell, K. E., Heyduk, K., Hibberd, J. M., Hodel, R. G. J., Hollingsworth, P. M., Johnson, M. T. J., Jost, R., Joyce, B., Kapralov, M. V., Kazamia, E., Kellogg, E. A., Koch, M. A., Von Konrat, M., Könyves, K., Kutchan, T. M., Lam, V., Larsson, A., Leitch, A. R., Lentz, R., Li, F.-W., Lowe, A. J., Ludwig, M., Manos, P. S., Mavrodiev, E., McCormick, M. K., McKain, M., McLellan, T., McNeal, J. R., Miller, R. E., Nelson, M. N., Peng, Y., Ralph, P., Real, D., Riggins, C. W., Ruhsam, M., Sage, R. F., Sakai, A. K., Scascitella, M., Schilling, E. E., Schlösser, E.-M., Sederoff, H., Servick, S., Sessa, E. B., Shaw, A. J., Shaw, S. W., Sigel, E. M., Skema, C., Smith, A. G., Smithson, A., Stewart, C. N., Stinchcombe, J. R., Szövényi, P., Tate, J. A., Tiebel, H., Trapnell, D., Villegente, M., Wang, C.-N., Weller, S. G., Wenzel, M., Weststrand, S., Westwood, J. H., Whigham, D. F., Wu, S., Wulff, A. S., Yang, Y., Zhu, D., Zhuang, C., Zuidof, J., Chase, M. W., Pires, J. C., Rothfels, C. J., Yu, J., Chen, C., Chen, L., Cheng, S., Li, J., Li, R., Li, X., Lu, H., Ou, Y., Sun, X., Tan, X., Tang, J., Tian, Z., Wang, F., Wang, J., Wei, X., Xu, X., Yan, Z., Yang, F., Zhong, X., Zhou, F., Zhu, Y., Zhang, Y., Ayyampalayam, S., Barkman, T. J., Nguyen, N.-p., Matasci, N., Nelson, D. R., Sayyari, E., Wafula, E. K., Walls, R. L., Warnow, T., An, H., Arrigo, N., Baniaga, A. E., Galuska, S., Jorgensen, S. A., Kidder, T. I., Kong, H., Lu-Irving, P., Marx, H. E., Qi, X., Reardon, C. R., Sutherland, B. L., Tiley, G. P., Welles, S. R., Yu, R., Zhan, S., Gramzow, L., Theißen, G., Wong, G. K.-S., and One Thousand Plant Transcriptomes Initiative (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, 574(7780):679–685.

Leigh, J. W., Schliep, K., Lopez, P., and Bapteste, E. (2011). Let them fall where they may: congruence analysis in massive phylogenetically messy data sets. *Molecular Biology and Evolution*, 28(10):2773–2785.

Leigh, J. W., Susko, E., Baumgartner, M., and Roger, A. J. (2008). Testing congruence in phylogenomic analysis. *Systematic Biology*, 57(1):104–115.

Lemmon, A. R. and Lemmon, E. M. (2012). High-throughput identification of informative nuclear loci for shallow-scale phylogenetics and phylogeography. *Systematic Biology*, 61(5):745–761.

Lemmon, E. M. and Lemmon, A. R. (2013). High-Throughput Genomic Data in Systematics and Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, 44(1):99–121.

Lemopoulos, A., Prokkola, J. M., Uusi-Heikkilä, S., Vasemägi, A., Huusko, A., Hyvärinen, P., Koljonen, M., Koskiniemi, J., and Vainikka, A. (2019). Comparing RADseq and microsatellites for estimating genetic diversity and relatedness — Implications for brown trout conservation. *Ecology and Evolution*, 9(4):2106–2120.

Leus, L., Van Laere, K., De Riek, J., and Van Huylenbroeck, J. (2018). Rose. In Van Huylenbroeck, J., editor, *Ornamental Crops*, volume 11, pages 719–767. Springer International Publishing, Cham.

Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G., and Webb, W. W. (2003). Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. *Science*, 299(5607):682–686.

Lewis, P. O. (2001). Phylogenetic systematics turns over a new leaf. *Trends in Ecology & Evolution*, 16(1):30–37.

Lewis, W. H. (1959). A monograph of the genus *Rosa* in North America. I. *R. Acicularis. Brittonia*, 11(1):1–24.

Lewis, W. H. (1965). Monograph of *Rosa* in North America. V. Subgenus *Hesperhodos. Annals of the Missouri Botanical Garden*, 52(2):99–113.

Lewis, W. H. (2016). Nomenclatural novelties in *Rosa* (Rosaceae) subgenus *Rosa* recognized in North America. *Novon: A Journal for Botanical Nomenclature*, 25(1):22–46.

Lewis, W. H. and Bayse, R. E. (1961). Analysis of nine crosses between diploid *Rosa* species. *Proceedings of the Society for Horticultural Science*, 78:573–579.

Lewis, W. H., Ertter, B., and Bruneau, A. (2014). 7. *Rosa* Linnaeus, Sp. Pl. 1: 491. 1753; Gen. Pl. ed. 5, 217. 1754. In *Flora of North America*, volume 9. Oxford University Press, New York, NY, USA.

Li, C., Chen, B., Xu, X., Li, D., and Dong, J. (2018). Simple sequence repeat markers associated/linked with agronomic traits, as core primers, are eminently suitable for DNA fingerprinting in Upland cotton. *Breeding Science*, 68(4):393–403.

Li, J., He, C., Guo, P., Zhang, P., and Liang, D. (2017). A workflow of massive identification and application of intron markers using snakes as a model. *Ecology and Evolution*, 7(23):10042–10055.

Li, M., Wunder, J., Bissoli, G., Scarponi, E., Gazzani, S., Barbaro, E., Saedler, H., and Varotto, C. (2008). Development of COS genes as universally amplifiable markers for phylogenetic reconstructions of closely related plant species. *Cladistics*, 24(5):727–745.

Li, S., Qu, X., Zhong, M., Jiang, X., Dong, X., Yi, T., Tang, K., Dai, S., and Hu, J.-Y. (2019). Characterization of the complete chloroplast genome of *Rosa chinensis* 'Old Blush' (Rosaceae), an important cultivated Chinese rose. *Acta Horticulturae*, 1232:119–124.

Lindley, J. (1820). *Rosarium Monographia: Or, A Botanical History of Roses. To Wich is Added, an Appendix, for the Use of Cultivators.* J. Ridgeway, London, UK.

Liorzou, M., Pernet, A., Li, S., Chastellier, A., Thouroude, T., Michel, G., Malécot, V., Gaillard, S., Briée, C., Foucher, F., Oghina-Pavie, C., Clotault, J., and Grapin, A. (2016). Nineteenth century French rose (*Rosa* sp.) germplasm shows a shift over time from a European to an Asian genetic background. *Journal of Experimental Botany*, 67(15):4711–4725.

Liston, A. (2014). 257 nuclear genes for Rosaceae phylogenomics.

Liu, C., Wang, G., Wang, H., Xia, T., Zhang, S., Wang, Q., and Fang, Y. (2015). Phylogenetic relationships in the genus *Rosa* revisited based on *rpl16*, *trnL-F*, and *atpB-rbcL* sequences. *HortScience*, 50(11):1618–1624.

Liu, M., Zhao, J., Wang, J., Liu, Z., and Liu, G. (2017). Phylogenetic analysis of 25 plant species representing 19 angiosperm families and one gymnosperm family based on 390 orthologous genes. *Plant Systematics and Evolution*, 303(3):413–417.

Liu, Z., Pagani, M., Zinniker, D., DeConto, R., Huber, M., Brinkhuis, H., Shah, S. R., Leckie, R. M., and Pearson, A. (2009). Global Cooling During the Eocene-Oligocene Climate Transition. *Science*, 323(5918):1187–1190.

Liò, P. and Goldman, N. (1998). Models of Molecular Evolution and Phylogeny. *Genome Research*, 8(12):1233–1244.

Lo, E. Y. Y., Stefanović, S., and Dickinson, T. A. (2010). Reconstructing reticulation history in a phylogenetic framework and the potential of allopatric speciation driven by polyploidy in an agamic complex in *Crataegus* (Rosaceae). *Evolution; International Journal of Organic Evolution*, 64(12):3593–3608.

Lyu, J., Song, J., Liu, Y., Wang, Y., Li, J., and Du, F. K. (2018). Species boundaries between three sympatric oak species: *Quercus aliena*, *Q. dentata*, and *Q. variabilis* at the Northern edge of their distribution in China. *Frontiers in Plant Science*, 9(414).

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: An *R* Package for Multivariate Analysis. *Journal of Statistical Software*, 25(1).

López-Giráldez, F. and Townsend, J. P. (2011). PhyDesign: an online application for profiling phylogenetic informativeness. *BMC Evolutionary Biology*, 11(152).

MacPhail, V. J. (2007). Pollination biology of wild roses (*Rosa* spp.) in Eastern Canada. Master's thesis, University of Guelph.

MacPhail, V. J. and Kevan, P. G. (2009). Review of the breeding systems of wild Roses (*Rosa* spp.). *Floriculture and Ornamental Biotechnology*, 3(1):1–13.

Maddison, W. P. and Wiens, J. J. (1997). Gene trees in species trees. *Systematic Biology*, 46(3):523–536.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2018). Cluster: Cluster Analysis Basics and Extensions.

Magoč, T. and Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21):2957–2963.

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends in Ecology & Evolution*, 20(5):229–237.

Malécot, V. (2015). Deux siècles de classification des roses sauvages et cultivées. In *Roses, mettez-vous au parfum Tome 1*, pages 25–28, Lyon. Société Nationale d'Horticulture de France.

Mandel, J. R., Dikow, R. B., Funk, V. A., Masalia, R. R., Staton, S. E., Kozik, A., Michelmore, R. W., Rieseberg, L. H., and Burke, J. M. (2014). A target enrichment method for gathering phylogenetic information from hundreds of loci: An example from the Compositae. *Applications in Plant Sciences*, 2(2).

Maraci, O., Ozkan, H., and Bilgin, R. (2018). Phylogeny and genetic structure in the genus *Secale*. *PLoS ONE*, 13(7):e0200825.

Marcussen, T., Heier, L., Brysting, A. K., Oxelman, B., and Jakobsen, K. S. (2015). From gene trees to a dated allopolyploid network: Insights from the angiosperm genus *Viola* (Violaceae). *Systematic Biology*, 64(1):84–101.

Marcussen, T., Jakobsen, K. S., Danihelka, J., Ballard, H. E., Blaxland, K., Brysting, A. K., and Oxelman, B. (2012). Inferring species networks from gene trees in high-polyploid North American and Hawaiian violets (*Viola*, Violaceae). *Systematic Biology*, 61(1):107–126.

Martin, W. F. and Cerff, R. (2017). Physiology, phylogeny, early evolution, and GAPDH. *Protoplasma*, 254(5):1823–1834.

Massatti, R., Reznicek, A. A., and Knowles, L. L. (2016). Utilizing RADseq data for phylogenetic analysis of challenging taxonomic groups: A case study in *Carex* sect. *Racemosae*. *American Journal of Botany*, 103(2):337–347.

Masure, P. (2013). *Guide des rosiers sauvages: 500 espèces, variétés et hybrides du monde.* Delachaux et Niestlé, Paris, France.

Matasci, N., Hung, L.-H., Yan, Z., Carpenter, E. J., Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Ayyampalayam, S., Barker, M., Burleigh, J. G., Gitzendanner, M. A., Wafula, E., Der, J. P., dePamphilis, C. W., Roure, B., Philippe, H., Ruhfel, B. R., Miles, N. W., Graham, S. W., Mathews, S., Surek, B., Melkonian, M., Soltis, D. E., Soltis, P. S., Rothfels, C., Pokorny, L., Shaw, J. A., DeGironimo, L., Stevenson, D. W., Villarreal, J. C., Chen, T., Kutchan, T. M., Rolf, M., Baucom, R. S., Deyholos, M. K., Samudrala, R., Tian, Z., Wu, X., Sun, X., Zhang, Y., Wang, J., Leebens-Mack, J., and Wong, G. K.-S. (2014). Data access for the 1,000 Plants (1kp) project. *GigaScience*, 3(17).

Matsumoto, S., Kouchi, M., Fukui, H., and Ueda, Y. (2000a). Phylogenetic analyses of the subgenus *Eurosa* using the ITS nrDNA sequence. *Acta Horticulturae*, 521:193–202.

Matsumoto, S., Kouchi, M., Yabuki, J., Kusunoki, M., Ueda, Y., and Fukui, H. (1998). Phylogenetic analyses of the genus *Rosa* using the *mat*K sequence: molecular evidence for the narrow genetic background of modern roses. *Scientia Horticulturae*, 77(1):73–82.

Matsumoto, S., Nishio, H., Ueda, Y., and Fukui, H. (2000b). Phylogenetic analyses of genus *Rosa*: polyphyly of section *Pimpinellifoliae* and origin of *Rosa* × *fortuniana* Lindl. *Acta Horticulturae*, 547:357–363.

Mayr, E. (1942). *Systematics and the origin of species, from the viewpoint of a zoologist.* Harvard University Press. Google-Books-ID: s2EGAAAAMAAJ.

Mayr, E. and Bock, W. J. (2002). Classifications and other ordering systems. *Journal of Zoological Systematics and Evolutionary Research*, 40(4):169–194.

McFadden, G. I. and van Dooren, G. G. (2004). Evolution: Red Algal Genome Affirms a Common Origin of All Plastids. *Current Biology*, 14(13):R514–R516.

McGregor Reid, G. (2009). Carolus Linnaeus (1707-1778): His Life, Philosophy and Science and Its Relationship to Modern Biology and Medicine. *Taxon*, 58(1):18–31.

McKain, M. R., Johnson, M. G., Uribe-Convers, S., Eaton, D., and Yang, Y. (2018). Practical considerations for plant phylogenomics. *Applications in Plant Sciences*, 6(3).

Meng, C. and Kubatko, L. S. (2009). Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. *Theoretical Population Biology*, 75(1):35–45.

Meng, G., Li, Y., Yang, C., and Liu, S. (2019). MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. *Nucleic Acids Research*, 47(11):e63–e63.

Meng, J., Fougère-Danezan, M., Zhang, L.-B., Li, D.-Z., and Yi, T.-S. (2011). Untangling the hybrid origin of the Chinese tea roses: evidence from DNA sequences of single-copy nuclear and chloroplast genes. *Plant Systematics and Evolution*, 297(3-4):157–170.

Mercier, D. (2014). *Rosa* L. In *Flora Gallica: Flore de France*, pages 996–1003. Biotope, Mèze, France.

Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1):31–46.

Michu, E. (2008). A short guide to phylogeny reconstruction. *Plant, Soil and Environment*, 53(10):442–446.

Millan, T., Osuna, F., Cobos, S., Torres, A. M., and Cubero, J. I. (1996). Using RAPDs to study phylogenetic relationships in *Rosa*. *TAG Theoretical and Applied Genetics*, 92(2):273–277.

Miller, J. H., Chambliss, E. B., and Bargeron, C. T. (2004). Invasive Plants of the Thirteen Southern States.

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., and Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, 17(2):240–248.

Minias, P., Dunn, P. O., Whittingham, L. A., Johnson, J. A., and Oyler-McCance, S. J. (2019). Evaluation of a Chicken 600k SNP genotyping array in non-model species of grouse. *Scientific Reports*, 9(6407).

Miyazato, P., Katsuya, H., Fukuda, A., Uchiyama, Y., Matsuo, M., Tokunaga, M., Hino, S., Nakao, M., and Satou, Y. (2016). Application of targeted enrichment to next-generation sequencing of retroviruses integrated into the host human genome. *Scientific Reports*, 6(28324).

Mudelsee, M., Bickert, T., Lear, C. H., and Lohmann, G. (2014). Cenozoic climate changes: A review based on time series analysis of marine benthic δ^{18}O records. *Reviews of Geophysics*, 52(3):333–374.

Nadeem, M. A., Nawaz, M. A., Shahid, M. Q., Doğan, Y., Comertpay, G., Yıldız, M., Hatipoğlu, R., Ahmad, F., Alsaleh, A., Labhane, N., Özkan, H., Chung, G., and Baloch, F. S. (2017).

DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. *Biotechnology & Biotechnological Equipment*.

Naidoo, K., Steenkamp, E. T., Coetzee, M. P. A., Wingfield, M. J., and Wingfield, B. D. (2013). Concerted Evolution in the Ribosomal RNA Cistron. *PLoS ONE*, 8(3).

Nakamura, N., Hirakawa, H., Sato, S., Otagaki, S., Matsumoto, S., Tabata, S., and Tanaka, Y. (2017). Genome structure of *Rosa multiflora*, a wild ancestor of cultivated roses. *DNA Research*, 25(2):113–121.

Nanjaraj Urs, A. N., Hu, Y., Li, P., Yuchi, Z., Chen, Y., and Zhang, Y. (2019). Cloning and Expression of a Nonribosomal Peptide Synthetase to Generate Blue Rose. *ACS synthetic biology*, 8(8):1698–1704.

Narechania, A., Baker, R., DeSalle, R., Mathema, B., Kolokotronis, S.-O., Kreiswirth, B., and Planet, P. J. (2016). Clusterflock: a flocking algorithm for isolating congruent phylogenomic datasets. *GigaScience*, 5(44).

Naumann, J., Symmank, L., Samain, M.-S., Müller, K. F., Neinhuis, C., Wanke, S., and others (2011). Chasing the hare - Evaluating the phylogenetic utility of a nuclear single copy gene region at and below species level within the species rich group *Peperomia* (Piperaceae). *BMC Evolutionary Biology*, 11(357).

Newberry, P. E. (1889). On the vegetable remains discovered in the cemetery at Hawara. In *WM Flinders Petrie. Hawara, Biahmu, and Arsinoe*, pages 46–53. The Leadenhall Press, London, Field & Tuer edition.

Nichols, R. (2001). Gene trees and species trees are not the same. *Trends in Ecology & Evolution*, 16(7):358–364.

Nikolov, L. A., Shushkov, P., Nevado, B., Gan, X., Al-Shehbaz, I. A., Filatov, D., Bailey, C. D., and Tsiantis, M. (2019). Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity. *New Phytologist*, 222(3):1638–1651.

Nishihara, H., Okada, N., and Hasegawa, M. (2007). Rooting the eutherian tree: the power and pitfalls of phylogenomics. *Genome Biology*, 8(9):R199.

Njuguna, W., Liston, A., Cronn, R., Ashman, T.-L., and Bassil, N. (2013). Insights into phylogeny, sex function and age of *Fragaria* based on whole chloroplast genome sequencing. *Molecular Phylogenetics and Evolution*, 66(1):17–29.

Nybom, H., Esselink, G. D., Werlemark, G., Leus, L., and Vosman, B. (2006). Unique genomic configuration revealed by microsatellite DNA in polyploid dogroses, *Rosa* sect. *Caninae. Journal of Evolutionary Biology*, 19(2):635–648.

Nybom, H., Esselink, G. D., Werlemark, G., and Vosman, B. (2004). Microsatellite DNA marker inheritance indicates preferential pairing between two highly homologous genomes in polyploid and hemisexual dog-roses, *Rosa* L. sect. *Caninae* DC. *Heredity*, 92(3):139–150.

Oghina-Pavie, C. (2015). Rose and Pear Breeding in Nineteenth-Century France: The Practice and Science of Diversity. In Phillips, D. and Kingsland, S., editors, *New Perspectives on the History of Life Sciences and Agriculture*, volume 40, pages 53–72. Springer International Publishing, Cham.

Oksanen, J., Blanchet, G. F., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., and Wagner, H. (2018). vegan: Community Ecology Package.

Page, R. D. M. and Holmes, E. C. (2009). *Molecular Evolution: A Phylogenetic Approach*. John Wiley & Sons.

Pante, E., Abdelkrim, J., Viricel, A., Gey, D., France, S. C., Boisselier, M. C., and Samadi, S. (2015). Use of RAD sequencing for delimiting species. *Heredity*, 114(5):450–459.

Paradis, E., Blomberg, S., Bolker, B., Brown, J., Claude, J., Cuong, H. S., Desper, R., Didier, G., Durand, B., Dutheil, J., Ewing, R. J., Gascuel, O., Heibl, C., Ives, A., Jones, B., Krah, F., Lawson, D., Lefort, V., Legendre, P., Lemon, J., McCloskey, R., Nylander, J., Opgen-Rhein, R., Popescu, A.-A., Royer-Carenzi, M., Schliep, K., Strimmer, K., and Vienne, D. d. (2017). ape: Analyses of Phylogenetics and Evolution.

Patané, J. S. L., Martins, J., and Setubal, J. C. (2018). Phylogenomics. In Setubal, J. C., Stoye, J., and Stadler, P. F., editors, *Comparative Genomics*, volume 1704, pages 103–187. Springer New York, New York, NY, USA.

Perrier, X. and Jacquemoud-Collet, J. P. (2006). DARwin software. http://darwin.cirad.fr/, accessed December 12, 2019.

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., and Baurain, D. (2011). Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. *PLoS Biology*, 9(3):e1000602.

Pilet-Nayel, M.-L., Moury, B., Caffier, V., Montarry, J., Kerlan, M.-C., Fournet, S., Durel, C.-E., and Delourme, R. (2017). Quantitative Resistance to Plant Pathogens in Pyramiding Strategies for Durable Crop Protection. *Frontiers in Plant Science*, 8(1838).

Planet, P. J. and Sarkar, I. N. (2005). mILD: a tool for constructing and analyzing matrices of pairwise phylogenetic character incongruence tests. *Bioinformatics*, 21(24):4423–4424.

Poczai, P. and Hyvönen, J. (2010). Nuclear ribosomal spacer regions in plant phylogenetics: problems and prospects. *Molecular Biology Reports*, 37(4):1897–1912.

Pond, S. L. K., Frost, S. D. W., and Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, 21(5):676–679.

Potter, D., Luby, J. J., and Harrison, R. E. (2000). Phylogenetic relationships among species of *Fragaria* (Rosaceae) Inferred from non-coding nuclear and chloroplast DNA sequences. *Systematic Botany*, 25(2):337–348.

Potter, J. (2011). *The Rose.* Atlantic Books, London, UK.

Prasad, A. B., Allard, M. W., NISC Comparative Sequencing Program, and Green, E. D. (2008). Confirming the phylogeny of mammals by use of large comparative sequence data sets. *Molecular Biology and Evolution*, 25(9):1795–1808.

Prathapan, K. D., Pethiyagoda, R., Bawa, K. S., Raven, P. H., Rajan, P. D., and 172 co-signatories from 35 countries (2018). When the cure kills—CBD limits biodiversity research. *Science*, 360(6396):1405–1406.

Primack, R. B. (1982). Ultraviolet patterns in flowers, or flowers as viewed by insects. *Arnoldia*, 42(3):139–146.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155(2):945–959.

Prothero, D. R. (1994). The Late Eocene-Oligocene Extinctions. *Annual Review Of Earth And Planetary Sciences*, 22:145–165.

Pécrix, Y., Rallo, G., Folzer, H., Cigna, M., Gudin, S., and Le Bris, M. (2011). Polyploidization mechanisms: temperature environment can induce diploid gamete formation in *Rosa* sp. *Journal of Experimental Botany*, 62(10):3587–3597.

Qingsong, Z. (2000). Uplift and Environmental Changes of the Tibetan Plateau. In *Mountain Geoecology and Sustainable Development of the Tibetan Plateau*, pages 19–45. Du Zheng, Qingsong Zhang, Shaohong Wu (Eds.), 1st edition.

Qiu, X., Zhang, H., Jian, H., Zhou, N., Yan, H., and Tang, K. (2013). Genetic relationships of wild roses, old garden roses, and modern roses based on internal transcribed spacers and *mat*K sequences. *HortScience*, 48(12):1445–1451.

Qiu, X., Zhang, H., Wang, Q., Jian, H., Yan, H., Zhang, T., Wang, J., and Tang, K. (2012). Phylogenetic relationships of wild roses in China based on nrDNA and *mat*K data. *Scientia Horticulturae*, 140:45–51.

Rabobank (2016). World Floriculture Map. `https://research.rabobank.com/far/en/sectors/regional-food-agri/world_floriculture_map_2016.html`, accessed December 12, 2019.

Rahman, M., Zafar, Y., and Paterson, A. H. (2009). Gossypium DNA Markers: Types, Numbers, and Uses. In Paterson, A. H., editor, *Genetics and Genomics of Cotton*, Plant Genetics and Genomics: Crops and Models, pages 101–139. Springer US, New York, NY.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology*, 67(5):901–904.

Ramsey, J. and Schemske, D. W. (2002). Neopolyploidy in Flowering Plants. *Annual Review of Ecology and Systematics*, 33(1):589–639.

Rani, S., Gupta, R. C., Kumari, S., and Rana, J. C. (2013). Chromosomal diversity in three species of genus *Rosa* L. (Rosaceae) from district Kangra of Himachal Pradesh. *Indian Journal of Genetics and Plant Breeding*, 73(4):454.

Raymond, O., Gouzy, J., Just, J., Badouin, H., Verdenaud, M., Lemainque, A., Vergne, P., Moja, S., Choisne, N., Pont, C., Carrère, S., Caissard, J.-C., Couloux, A., Cottret, L., Aury, J.-M., Szécsi, J., Latrasse, D., Madoui, M.-A., François, L., Fu, X., Yang, S.-H., Dubois, A., Piola, F., Larrieu, A., Perez, M., Labadie, K., Perrier, L., Govetto, B., Labrousse, Y., Villand, P., Bardoux, C., Boltz, V., Lopez-Roques, C., Heitzler, P., Vernoux, T., Vandenbussche, M., Quesneville, H., Boualem, A., Bendahmane, A., Liu, C., Bris, M. L., Salse, J., Baudino, S., Benhamed, M., Wincker, P., and Bendahmane, M. (2018). The *Rosa* genome provides new insights into the domestication of modern roses. *Nature Genetics*, 50(6):772–777.

Reboud, X. and Zeyl, C. (1994). Organelle inheritance in plants. *Heredity*, 72(2):132–140.

Redouté, P. J. and Thory, C. A. (1817). *Les Roses.* Firmin-Didot, Paris, France.

Refulio-Rodriguez, N. F. and Olmstead, R. G. (2014). Phylogeny of Lamiidae. *American Journal of Botany*, 101(2):287–299.

Rehder, A. (1940). *Rosa* L. In *Manual of cultivated trees and shrubs hardy in North America*, pages 426–452. Collier MacMillan Ltd, New York, NY, USA.

Remay, A., Lalanne, D., Thouroude, T., Couviour, F. L., Oyant, L. H.-S., and Foucher, F. (2009). A survey of flowering genes reveals the role of gibberellins in floral control in rose. *Theoretical and Applied Genetics*, 119(5):767–781.

Ren, G., Mateo, R. G., Liu, J., Suchan, T., Alvarez, N., Guisan, A., Conti, E., and Salamin, N. (2017). Genetic consequences of Quaternary climatic oscillations in the Himalayas: *Primula tibetica* as a case study based on restriction site-associated DNA sequencing. *The New Phytologist*, 213(3):1500–1512.

Ren, R., Wang, H., Guo, C., Zhang, N., Zeng, L., Chen, Y., Ma, H., and Qi, J. (2018). Widespread whole genome duplications contribute to genome complexity and species diversity in Angiosperms. *Molecular Plant*, 11(3):414–428.

Renny-Byfield, S., Chester, M., Kovařík, A., Le Comber, S. C., Grandbastien, M.-A., Deloger, M., Nichols, R. A., Macas, J., Novák, P., Chase, M. W., and Leitch, A. R. (2011). Next Generation Sequencing reveals genome downsizing in allotetraploid *Nicotiana tabacum*, Predominantly through the elimination of paternally derived repetitive DNAs. *Molecular Biology and Evolution*, 28(10):2843–2854.

Reuter, J., Spacek, D. V., and Snyder, M. (2015). High-throughput sequencing technologies. *Molecular Cell*, 58(4):586–597.

Ritz, C. M., Schmuths, H., and Wissemann, V. (2005). Evolution by reticulation: European dogroses originated by multiple hybridization across the genus *Rosa*. *Journal of Heredity*, 96(1):4–14.

Ritz, C. M. and Wissemann, V. (2011). Microsatellite analyses of artificial and spontaneous dogrose hybrids reveal the hybridogenic origin of *Rosa micrantha* by the contribution of unreduced gametes. *Journal of Heredity*, 102(2):217–227.

Roberts, A. V. (1977). Relationship between species in the genus *Rosa*, section *Pimpinellifoliae*. *Botanical Journal of the Linnean Society*, 74(4):309–328.

Roberts, A. V., Gladis, T., and Brumme, H. (2009). DNA amounts of roses (*Rosa* L.) and their use in attributing ploidy levels. *Plant Cell Reports*, 28(1):61–71.

Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1):131–147.

Robyns, W. (1938). G. A. Boulenger 1858-1937. Sa vie et son oeuvre rhodologique. *Bulletin du Jardin botanique de l'État à Bruxelles*, 15(1):1–24.

Rothfels, C. J., Pryer, K. M., and Li, F.-W. (2017). Next-generation polyploid phylogenetics: rapid resolution of hybrid polyploid complexes using PacBio single-molecule sequencing. *New Phytologist*, 213(1):413–429.

Roure, B., Rodriguez-Ezpeleta, N., and Philippe, H. (2007). SCaFoS: a tool for Selection, Concatenation and Fusion of Sequences for phylogenomics. *BMC Evolutionary Biology*, 7 (Suppl 1)(S2).

Rousseau-Gueutin, M., Keller, J., Ferreira de Carvalho, J., Aïnouche, A., and Martin, G. (2018). The Intertwined Chloroplast and Nuclear Genome Coevolution in Plants. In Ratnadewi, D. and Hamim, editors, *Plant Growth and Regulation - Alterations to Sustain Unfavorable Conditions*. IntechOpen.

Rowley, G. (1960). Aneuploidy in the genus *Rosa*. *Journal of Genetics*, 57(2-3):253–268.

Rowley, G. D. (1955). Hulthemosas - New Hope for the Rose Breeder. *The Rose Annual*, pages 37–40.

Ruhfel, B. R., Gitzendanner, M. A., Soltis, P. S., Soltis, D. E., and Burleigh, J. G. (2014). From algae to angiosperms – inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evolutionary Biology*, 14(23).

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425.

Salichos, L. and Rokas, A. (2013). Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, 497(7449):327–331.

Salinas, N. R. and Little, D. P. (2014). 2matrix: A Utility for Indel Coding and Phylogenetic Matrix Concatenation. *Applications in Plant Sciences*, 2(1):1300083.

Sang, T. (2002). Utility of low-copy nuclear gene sequences in plant phylogenetics. *Critical Reviews in Biochemistry and Molecular Biology*, 37(3):121–147.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467.

Sanjur, O. I., Piperno, D. R., Andres, T. C., and Wessel-Beaver, L. (2002). Phylogenetic relationships among domesticated and wild species of *Cucurbita* (Cucurbitaceae) inferred from a mitochondrial gene: Implications for crop plant evolution and areas of origin. *Proceedings of the National Academy of Sciences of the United States of America*, 99(1):535–540.

Sarver, B. A., Keeble, S., Cosart, T., Tucker, P. K., Dean, M. D., and Good, J. M. (2017). Phylogenomic Insights into Mouse Evolution Using a Pseudoreference Approach. *Genome Biology and Evolution*, 9(3):726–739.

Scalliet, G., Piola, F., Douady, C. J., Réty, S., Raymond, O., Baudino, S., Bordji, K., Bendahmane, M., Dumas, C., Cock, J. M., and Hugueney, P. (2008). Scent evolution in Chinese roses. *Proceedings of the National Academy of Sciences*, 105(15):5927–5932.

Scariot, V., Akkak, A., and Botta, R. (2006). Characterization and genetic relationships of wild species and old garden roses based on microsatellite analysis. *Journal of the American Society for Horticultural Science*, 131(1):66–73.

Schori, M. and Showalter, A. M. (2011). DNA barcoding as a means for identifying medicinal plants of Pakistan. *Pakistan Journal of Botany*, 43:1–4.

Schramm, D. g. (2016). Damask roses: an untold story. *Rose letter*, 40(2):2–7.

Seringe, N. C. (1823). *Musée Helvétique d'Histoire Naturelle. (Partie botanique.) ou collection de mémoires, monographies, notices botaniques avec 16 planches sur cuivre ou lithographiées*, volume 1. Imprimerie de L. Alb. Haller, Genève & Berne.

Shaw, J., Lickey, E. B., Beck, J. T., Farmer, S. B., Liu, W., Miller, J., Siripun, K. C., Winder, C. T., Schilling, E. E., and Small, R. L. (2005). The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany*, 92(1):142–166.

Shaw, J., Lickey, E. B., Schilling, E. E., and Small, R. L. (2007). Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *American Journal of Botany*, 94(3):275–288.

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., Jaiswal, P., Mockaitis, K., Liston, A., Mane, S. P., Burns, P., Davis, T. M., Slovin, J. P., Bassil, N., Hellens, R. P., Evans, C., Harkins, T., Kodira, C., Desany, B., Crasta, O. R., Jensen, R. V., Allan, A. C., Michael, T. P., Setubal, J. C., Celton, J.-M., Rees, D. J. G., Williams, K. P., Holt, S. H., Rojas, J. J. R., Chatterjee, M., Liu, B., Silva, H., Meisel, L., Adato, A., Filichkin, S. A., Troggio, M., Viola, R., Ashman, T.-L., Wang, H., Dharmawardhana, P., Elser, J., Raja, R., Priest, H. D., Bryant Jr, D. W., Fox, S. E., Givan, S. A., Wilhelm, L. J., Naithani, S., Christoffels, A., Salama, D. Y., Carter, J., Girona, E. L., Zdepski, A., Wang, W., Kerstetter, R. A., Schwab, W., Korban, S. S., Davik, J., Monfort, A., Denoyes-Rothan, B., Arus, P., Mittler, R., Flinn, B., Aharoni, A., Bennetzen, J. L., Salzberg, S. L., Dickerman, A. W., Velasco, R., Borodovsky, M., Veilleux, R. E., and Folta, K. M. (2011). The genome of woodland strawberry (*Fragaria vesca*). *Nature Genetics*, 43(2):109–116.

Sieber, J. (2009). ADR-Chronik. Technical report, Allgemeine Deutsche Rosenneuheitenprüfung.

Singh, S., Dhyani, D., Nag, A., and Sharma, R. K. (2017). Morphological and molecular characterization revealed high species level diversity among cultivated, introduced and wild roses (*Rosa* sp.) of western Himalayan region. *Genetic Resources and Crop Evolution*, 64(3):515–530.

Sloan, P. R. (1976). The Buffon-Linnaeus Controversy. *Isis*, 67(3):356–375.

Small, R. L., Cronn, R. C., and Wendel, J. F. (2004). L. A. S. JOHNSON REVIEW No. 2. Use of nuclear genes for phylogeny reconstruction in plants. *Australian Systematic Botany*, 17(2):145–170.

Smith, L. M., Fung, S., Hunkapiller, M. W., Hunkapiller, T. J., and Hood, L. E. (1985). The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic Acids Research*, 13(7):2399–2412.

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B. H., and Hood, L. E. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature*, 321(6071):674–679.

Smith, S. A., Moore, M. J., Brown, J. W., and Yang, Y. (2015). Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology*, 15(150).

Smulders, M. J. M., Arens, P., Bourke, P. M., Debener, T., Linde, M., Riek, J. D., Leus, L., Ruttink, T., Baudino, S., Saint-Oyant, L. H., Clotault, J., and Foucher, F. (2019). In the name of the rose: a roadmap for rose research in the genome era. *Horticulture Research*, 6(65).

Smulders, M. J. M., Arens, P., Koning-Boucoiran, C. F. S., Gitonga, V. W., Krens, F. A., Atanassov, A., Atanassov, I., Rusanov, K. E., Bendahmane, M., Dubois, A., Raymond, O., Caissard, J. C., Baudino, S., Crespel, L., Gudin, S., Ricci, S. C., Kovatcheva, N., Van Huylenbroeck, J., Leus, L., Wissemann, V., Zimmermann, H., Hensen, I., Werlemark, G., and Nybom, H. (2011). *Rosa*. In Kole, C., editor, *Wild Crop Relatives: Genomic and Breeding Resources*, pages 243–275. Springer Berlin Heidelberg, Berlin, Heidelberg.

Smýkal, P., Kenicer, G., Flavell, A. J., Corander, J., Kosterin, O., Redden, R. J., Ford, R., Coyne, C. J., Maxted, N., Ambrose, M. J., and Ellis, N. T. H. (2011). Phylogeny, phylogeography and genetic diversity of the *Pisum* genus. *Plant Genetic Resources*, 9(1):4–18.

Sokal, R. R. (1986). Phenetic Taxonomy: Theory and Methods. *Annual Review of Ecology and Systematics*, 17:423–442.

Sokal, R. R. and Sneath, P. H. A. (1963). *Principles of numerical taxonomy.* W. H. Freeman, San Franciso, CA, USA.

Soltis, D. E., Smith, S. A., Cellinese, N., Wurdack, K. J., Tank, D. C., Brockington, S. F., Refulio-Rodriguez, N. F., Walker, J. B., Moore, M. J., Carlsward, B. S., Bell, C. D., Latvis, M., Crawley, S., Black, C., Diouf, D., Xi, Z., Rushworth, C. A., Gitzendanner, M. A., Sytsma, K. J., Qiu, Y.-L., Hilu, K. W., Davis, C. C., Sanderson, M. J., Beaman, R. S., Olmstead, R. G., Judd, W. S., Donoghue, M. J., and Soltis, P. S. (2011). Angiosperm phylogeny: 17 genes, 640 taxa. *American Journal of Botany*, 98(4):704–730.

Soltis, P. S. and Soltis, D. E. (2009). The role of hybridization in plant speciation. *Annual Review of Plant Biology*, 60:561–588.

Solís-Lemus, C., Bastide, P., and Ané, C. (2017). PhyloNetworks: A Package for Phylogenetic Networks. *Molecular Biology and Evolution*, 34(12):3292–3298.

Sosef, M. S. M. (1997). Hierarchical Models, Reticulate Evolution and the Inevitability of Paraphyletic Supraspecific Taxa. *Taxon*, 46(1):75–85.

Spooner, D. M. (2016). Species delimitations in plants: lessons learned from potato taxonomy by a practicing taxonomist. *Journal of Systematics and Evolution*, 54(3):191–203.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

Stebbins, G. L. (1950). *Variation and Evolution in Plants.* Columbia University Press.

Stevens, P. F. (2017). Angiosperm Phylogeny Website, version 14, July 2017 [and more or less continuously updated since].

Stoffelen, P., Verdegem, I., Hoste, I., Diagre, D., Janssens, S., De Smedt, S., Hanquart, N., Groom, Q., Bogaerts, A., Mergen, P., and de Briey, H. (2018). Opening-up Crépin's Rose Herbarium by New Technologies: a Pilot Project. *Biodiversity Information Science and Standards*, 2:e25792.

Straub, S. C. K., Parks, M., Weitemier, K., Fishbein, M., Cronn, R. C., and Liston, A. (2012). Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany*, 99(2):349–364.

Su, T., Huang, Y.-J., Meng, J., Zhang, S.-T., Huang, J., and Zhou, Z.-K. (2016). A Miocene leaf fossil record of *Rosa* (*R. fortuita* n. sp.) from its modern diversity center in SW China. *Palaeoworld*, 25(1):104–115.

Su, T., Jacques, F. M. B., Spicer, R. A., Liu, Y.-S., Huang, Y.-J., Xing, Y.-W., and Zhou, Z.-K. (2013). Post-Pliocene establishment of the present monsoonal climate in SW China: evidence from the late Pliocene Longmen megaflora. *Climate of the Past*, 9(4):1911–1920.

Sun, J., Ni, X., Bi, S., Wu, W., Ye, J., Meng, J., and Windley, B. F. (2014). Synchronous turnover of flora, fauna, and climate at the Eocene–Oligocene Boundary in Asia. *Scientific Reports*, 4(7463).

Sánchez-Baracaldo, P., Raven, J. A., Pisani, D., and Knoll, A. H. (2017). Early photosynthetic eukaryotes inhabited low-salinity habitats. *Proceedings of the National Academy of Sciences of the United States of America*, 114(37):E7737–E7745.

Tabor, P. (1960). The Cherokee Rose. *The Georgia Review*, 14(3):231–236.

Tamuri, A. and Goldman, N. (2017). Avoiding ascertainment bias in the maximum likelihood inference of phylogenies based on truncated data. *bioRχiv*, 186478.

Tan, J., Wang, J., Luo, L., Yu, C., Xu, T., Wu, Y., Cheng, T., Wang, J., Pan, H., and Zhang, Q. (2017). Genetic relationships and evolution of old Chinese garden roses based on SSRs and chromosome diversity. *Scientific Reports*, 7(15437).

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17:57–86.

Than, C., Ruths, D., and Nakhleh, L. (2008). PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, 9(322).

Thulin, M. (1993). *Flora of Somalia*, volume 1. Royal Botanic Gardens, Kew, UK.

Tiffney, B. H. and Manchester, S. R. (2001). The Use of Geological and Paleontological Evidence in Evaluating Plant Phylogeographic Hypotheses in the Northern Hemisphere Tertiary. *International Journal of Plant Sciences*, 162(S6):S3–S17.

Tomasello, S. (2018). How many names for a beloved genus? – Coalescent-based species delimitation in *Xanthium* L. (Ambrosiinae, Asteraceae). *Molecular Phylogenetics and Evolution*, 127:135–145.

Tomljenovic, N. and Pejic, I. (2018). Taxonomic review of the genus *Rosa*. *Agriculturae Conspectus Scientificus*, 83(2):139–147.

Touw, M. (1982). Roses in the Middle Ages. *Economic Botany*, 36(1):71–83.

Townsend, J. P. (2007). Profiling phylogenetic informativeness. *Systematic Biology*, 56(2):222–231.

Townsend, J. P. and Leuenberger, C. (2011). Taxon sampling and the optimal rates of evolution for phylogenetic inference. *Systematic Biology*, 60(3):358–365.

Tremetsberger, K., Ortiz, M. A., Terrab, A., Balao, F., Casimiro-Soriguer, R., Talavera, M., and Talavera, S. (2016). Phylogeography above the species level for perennial species in a composite genus. *AoB PLANTS*, 8.

Tucker, A. O. (2004). Identification of the rose, sage, iris, and lily in the “Blue Bird Fresco” from Knossos, Crete (ca. 1450 B.C.E.). *Economic Botany*, 58(4):733–736.

Turland, N., Wiersema, J., Barrie, F., Greuter, W., Hawksworth, D., Herendeen, P., Knapp, S., Kusber, W.-H., Li, D.-Z., Marhold, K., May, T., McNeill, J., Monro, A., Prado, J., Price, M., and Smith, G., editors (2018). *International Code of Nomenclature for algae, fungi, and plants*, volume 159 of *Regnum Vegetabile*. Koeltz Botanical Books.

Täckholm, G. (1920). On the cytology of the genus *Rosa*. *Svensk Botanisk Tidskrift*, 14:300–311.

Ungerer, M. C., Strakosh, S. C., and Zhen, Y. (2006). Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. *Current biology: CB*, 16(20):R872–873.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., and Rozen, S. G. (2012). Primer3 - new capabilities and interfaces. *Nucleic Acids Research*, 40(15).

Uribe-Convers, S., Settles, M. L., and Tank, D. C. (2016). A Phylogenomic Approach Based on PCR Target Enrichment and High Throughput Sequencing: Resolving the Diversity within the South American Species of *Bartsia* L. (Orobanchaceae). *PLOS ONE*, 11(2):e0148203.

Vaezi, J., Arjmandi, A. A., and Sharghi, H. R. (2019). Origin of *Rosa* × *binaloudensis* (Rosaceae), a new natural hybrid species from Iran. *Phytotaxa*, 411(1):23–38.

VAL’HOR (2018). Roses et rosiers achétés par les Français en 2018. Technical report, Val’hor & France AgriMer.

van Dongen, S. (2012). mclblastline - a pipeline for clustering from BLAST files.

van Zeist, W. and Palfenier-Vegter, R. M. (1981). Seeds and fruits from the Swifterbant S3 site. Final reports on Swifterbant IV. *Palaeohistoria*, 23:105–158.

VanBuren, R., Bryant, D., Bushakra, J. M., Vining, K. J., Edger, P. P., Rowley, E. R., Priest, H. D., Michael, T. P., Lyons, E., Filichkin, S. A., Dossett, M., Finn, C. E., Bassil, N. V., and Mockler, T. C. (2016). The genome of black raspberry (*Rubus occidentalis*). *The Plant Journal*, 87(6):535–547.

Verma, R. S., Chandra Padalia, R., and Chauhan, A. (2016). Rose-scented Geranium (*Pelargonium* sp.) oils. In *Essential Oils in Food Preservation, Flavor and Safety*, pages 697–704. Elsevier.

Vieira, M. L. C., Santini, L., Diniz, A. L., and Munhoz, C. d. F. (2016). Microsatellite markers: what they mean and why they are so useful. *Genetics and Molecular Biology*, 39(3):312–328.

Von Haeseler, A. (2012). Do we still need supertrees? *BMC biology*, 10(13).

Waddell, P. J., Kishino, H., and Ota, R. (2000). Rapid evaluation of the phylogenetic congruence of sequence data using likelihood ratio tests. *Molecular Biology and Evolution*, 17(12):1988–1992.

Walther, H. and Kvaček, Z. (2007). Early Oligocene flora of Seifhennersdorf (Saxony). *Acta Mus Nat Prag, Ser B, Hist Natur*, 63:85–174.

Wang, H.-J., Li, W.-T., Liu, Y.-N., Yang, F.-S., and Wang, X.-Q. (2017a). Resolving interspecific relationships within evolutionarily young lineages using RNA-seq data: An example from Pedicularis section Cyathophora (Orobanchaceae). *Molecular Phylogenetics and Evolution*, 107:345–355.

Wang, M., Zhang, C., Li, M., and Gao, X. (2019). The complete chloroplast genome sequence of *Rosa banksiae* var. *normalis* (Rosaceae). *Mitochondrial DNA Part B*, 4(1):969–970.

Wang, Q., Hu, H., An, J., Bai, G., Ren, Q., and Liu, J. (2018). Complete chloroplast genome sequence of *Rosa roxburghii* and its phylogenetic analysis. *Mitochondrial DNA Part B*, 3(1):149–150.

Wang, Y., Chen, Q., Chen, T., Tang, H., Liu, L., and Wang, X. (2016). Phylogenetic insights into Chinese *Rubus* (Rosaceae) from multiple chloroplast and nuclear DNAs. *Frontiers in Plant Science*, 7.

Wang, Y.-H., Comes, H. P., Cao, Y.-N., Guo, R., Mao, Y.-R., and Qiu, Y.-X. (2017b). Quaternary climate change drives allo-peripatric speciation and refugial divergence in the *Dysosma versipellis-pleiantha* complex from different forest types in China. *Scientific Reports*, 7(1):1–13.

Watts, M. T. (2009). *Reading the Landscape of Europe*. Nature Study Guild Publishers.

Welzen, P. C. v. (1997). Paraphyletic Groups or What Should a Classification Entail. *Taxon*, 46(1):99–103.

Wen, D., Yu, Y., Zhu, J., and Nakhleh, L. (2018). Inferring Phylogenetic Networks Using PhyloNet. *Systematic Biology*, 67(4):735–740.

Wen, J., Nie, Z.-L., and Ickert-Bond, S. M. (2016). Intercontinental disjunctions between eastern Asia and western North America in vascular plants highlight the biogeographic importance of the Bering land bridge from late Cretaceous to Neogene. *Journal of Systematics and Evolution*, 54(5):469–490.

Wendel, J. F., Schnabel, A., and Seelanan, T. (1995). Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). *Proceedings of the National Academy of Sciences of the United States of America*, 92(1):280–284.

Werlemark, G. and Nybom, H. (2017). Inheritance in the Dogrose. In *Reference Module in Life Sciences*. Elsevier.

Wetterstrand, K. A. (2019). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).

Whatley, J. M., John, P., and Whatley, F. R. (1979). From extracellular to intracellular: the establishment of mitochondria and chloroplasts. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 204(1155):165–187.

Wicke, S., Schneeweiss, G. M., dePamphilis, C. W., Müller, K. F., and Quandt, D. (2011). The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Molecular Biology*, 76(3-5):273–297.

Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M. S., Burleigh, J. G., Gitzendanner, M. A., Ruhfel, B. R., Wafula, E., Der, J. P., Graham, S. W., Mathews, S., Melkonian, M., Soltis, D. E., Soltis, P. S., Miles, N. W., Rothfels, C. J., Pokorny, L., Shaw, A. J., DeGironimo, L., Stevenson, D. W., Surek, B., Villarreal, J. C., Roure, B., Philippe, H., dePamphilis, C. W., Chen, T., Deyholos, M. K., Baucom, R. S., Kutchan, T. M., Augustin, M. M., Wang, J., Zhang, Y., Tian, Z., Yan, Z., Wu, X., Sun, X., Wong, G. K.-S., and Leebens-Mack, J. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences*, 111(45):E4859–E4868.

Widrlechner, M. P. (1981). History and Utilization of *Rosa damascena*. *Economic Botany*, 35(1):42–58.

Wiens, J. J. and Penkrot, T. A. (2002). Delimiting species using DNA and morphological variation and discordant species limits in spiny lizards (*Sceloporus*). *Systematic Biology*, 51(1):69–91.

Wissemann, V. (1999). Genetic constitution of *Rosa* sect. *Caninae* (*R. canina*, *R. jundzillii*) and sect. *Gallicanae* (*R. gallica*). *Journal of Applied Botany*, 73:191–196.

Wissemann, V. (2002). Molecular evidence for allopolyploid origin of the *Rosa canina*-complex (Rosaceae, Rosoideae). *Journal of Applied Botany/Angewandte Botanik*, 76:176–178.

Wissemann, V. (2003a). Conventional taxonomy (wild roses). In *Encyclopedia of Rose science*, pages 111–117. A.V. Roberts, T. Debener and S. Gudin (Eds.), Oxford, Elsevier edition.

Wissemann, V. (2003b). Hybridization and the evolution of the nrITS spacer region. In *Plant genome, biodiversity and evolution*, volume 1 Part A, pages 57–71. Science Publishers, Enfield, New Hampshire.

Wissemann, V. (2010). Plant evolution by means of hybridization. *Systematics and Biodiversity.*

Wissemann, V. and Ritz, C. M. (2005). The genus *Rosa* (Rosoideae, Rosaceae) revisited: molecular analysis of nrITS-1 and *atp*B-*rbc*L intergenic spacer (IGS) versus conventional taxonomy. *Botanical Journal of the Linnean Society*, 147(3):275–290.

Wissemann, V. and Ritz, C. M. (2007). Evolutionary patterns and processes in the genus *Rosa* (Rosaceae) and their implications for host-parasite co-evolution. *Plant Systematics and Evolution*, 266(1-2):79–89.

Wolfe, K. H., Li, W. H., and Sharp, P. M. (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences of the United States of America*, 84(24):9054–9058.

Wood, T. E., Takebayashi, N., Barker, M. S., Mayrose, I., Greenspoon, P. B., and Rieseberg, L. H. (2009). The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences*, 106(33):13875–13879.

Wortley, A. H., Rudall, P. J., Harris, D. J., and Scotland, R. W. (2005). How Much Data are Needed to Resolve a Difficult Phylogeny? Case Study in Lamiales. *Systematic Biology*, 54(5):697–709.

Wu, S., Ueda, Y., Nishihara, S., and Matsumoto, S. (2001). Phylogenetic analysis of Japanese *Rosa* species using DNA sequences of nuclear ribosomal internal trancribed spacers (ITS). *The Journal of Horticultural Science and Biotechnology*, 76(2):127–132.

Wuyts, S. and Segata, N. (2019). At the forefront of the sequencing revolution—notes from the RNGS19 conference. *Genome Biology*, 20(1):93.

Wylie, A. (1954). The History of Garden Roses, part 1. *Journal of the Royal Horticultural Society*, 79:555–571.

Xiang, Y., Huang, C.-H., Hu, Y., Wen, J., Li, S., Yi, T., Chen, H., Xiang, J., and Ma, H. (2016). Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Molecular Biology and Evolution*, 34(2):262–281.

Xu, X.-W., Wu, J.-W., Qi, M.-X., Lu, Q.-X., Lee, P. F., Lutz, S., Ge, S., and Wen, J. (2015). Comparative phylogeography of the wild-rice genus *Zizania* (Poaceae) in eastern Asia and North America. *American Journal of Botany*, 102(2):239–247.

Yang, J.-B., Li, D.-Z., and Li, H.-T. (2014). Highly effective sequencing whole chloroplast genomes of angiosperms by nine novel universal primer pairs. *Molecular Ecology Resources*, 14(5):1024–1031.

Yang, Y. and Davis, T. M. (2017). A New Perspective on Polyploid Fragaria (Strawberry) Genome Composition Based on Large-Scale, Multi-Locus Phylogenetic Analysis. *Genome Biology and Evolution*, 9(12):3433–3448.

Yang, Z. and Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, 13(5):303–314.

Yokoya, K., Roberts, A. V., Mottley, J., Lewis, R., and Brandham, P. E. (2000). Nuclear DNA amounts in roses. *Annals of Botany*, 85(4):557–561.

You, Q., Yang, X., Peng, Z., Xu, L., and Wang, J. (2018). Development and Applications of a High Throughput Genotyping Tool for Polyploid Crops: Single Nucleotide Polymorphism (SNP) Array. *Frontiers in Plant Science*, 9(104).

Younis, A., Ramzan, F., Hwang, Y.-J., and Lim, K.-B. (2015). FISH and GISH: molecular cytogenetic tools and their applications in ornamental plants. *Plant Cell Reports*, 34(9):1477–1488.

Yu, Y., Dong, J., Liu, K. J., and Nakhleh, L. (2014). Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences*, 111(46):16448–16453.

Yu, Y. and Nakhleh, L. (2015). A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*, 16(10):S10.

Yu, Y., Xiang, Q., Manos, P. S., Soltis, D. E., Soltis, P. S., Song, B.-H., Cheng, S., Liu, X., and Wong, G. (2017). Whole-genome duplication and molecular evolution in *Cornus* L. (Cornaceae) – Insights from transcriptome sequences. *PLoS ONE*, 12(2).

Zanella, C. M., Palma-Silva, C., Goetze, M., and Bered, F. (2016). Hybridization between two sister species of Bromeliaceae: Vriesea carinata and V. incurvata. *Botanical Journal of the Linnean Society*, 181(3):491–504.

Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018a). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, 19(Suppl 6)(153).

Zhang, J., Esselink, G. D., Che, D., Fougère-Danezan, M., Arens, P., and Smulders, M. J. M. (2013). The diploid origins of allopolyploid rose species studied using single nucleotide polymorphism haplotypes flanking a microsatellite repeat. *The Journal of Horticultural Science and Biotechnology*, 88(1):85–92.

Zhang, L., Zeng, T., Hu, H., Fan, L., Zheng, H., and Hu, Q. (2018b). Interspecific divergence of two *Sinalliaria* (Brassicaceae) species in Eastern China. *Frontiers in Plant Science*, 9.

Zhang, S.-D., Jin, J.-J., Chen, S.-Y., Chase, M. W., Soltis, D. E., Li, H.-T., Yang, J.-B., Li, D.-Z., and Yi, T.-S. (2017). Diversification of Rosaceae since the Late Cretaceous based on plastid phylogenomics. *New Phytologist*, 214(3):1355–1367.

Zhang, S.-D., Zhang, C., and Ling, L.-Z. (2019). The complete chloroplast genome of *Rosa berberifolia*. *Mitochondrial DNA Part B*, 4(1):1741–1742.

Zhao, K.-K., Landrein, S., Barrett, R. L., Sakaguchi, S., Maki, M., Mu, W.-X., Yang, T., Zhu, Z.-X., Liu, H., and Wang, H.-F. (2019). Phylogeographic Analysis and Genetic Structure of an Endemic Sino-Japanese Disjunctive Genus *Diabelia* (Caprifoliaceae). *Frontiers in Plant Science*, 10.

Zheng, Y.-H., Alverson, A. J., Wang, Q.-F., and Palmer, J. D. (2013). Chloroplast phylogeny of *Cucurbita*: Evolution of the domesticated and wild species. *Journal of Systematics and Evolution*, 51(3):326–334.

Zhu, Z.-M., Gao, X.-F., and Fougère-Danezan, M. (2015). Phylogeny of *Rosa* sections *Chinenses* and *Synstylae* (Rosaceae) based on chloroplast and nuclear markers. *Molecular Phylogenetics and Evolution*, 87:50–64.

Zlesak, D. C. (2009). Pollen diameter and guard cell length as predictors of ploidy in diverse rose cultivars, species, and breeding lines. *Floriculture and Ornamental Biotechnology*, 3(1):53–70.

Zwemer, S. M. and Zwemer, M. C. (1941). The Rose and Islam. *The Muslim World*, 31(4):360–370.

Zwickl, D. J. (2006). *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion.* PhD thesis, School of Biological Sciences, University of Texas at Austin, Austin, TX, USA.

Šarhanová, P., Pfanzelt, S., Brandt, R., Himmelbach, A., and Blattner, F. R. (2018). SSR-seq: Genotyping of microsatellites using next-generation sequencing reveals higher level of polymorphism as compared to traditional fragment size scoring. *Ecology and Evolution*, 8(22):10817–10833.

UNIVERSITE BRETAGNE LOIRE / ECOLOGIE GEOSCIENCES AGRONOMIE ALIMENTATION

université angers

Titre : Phylogénomique du genre *Rosa*: hybridation et polyploïdie comme facteurs de diversification

Mots clés : Phylogénomique; *Rosa* sp.; séquençage d'amplicons; analyses d'orthologie; réseaux

Résumé : Le genre *Rosa* comprend 150-200 espèces bien réparties sur l'hémisphère nord et présente une histoire évolutive complexe. L'hybridation et la polyploïdie sont des forces évolutives majeures chez *Rosa* bien que ces deux processus ont à peine été pris en compte dans les dernières phylogénies. Avec la récente acquisition de génomes complets et le développement de techniques de séquençage à haut débit, l'objectif de cette thèse était de développer un cadre phylogénomique général pour résoudre les relations phylogénétiques au sein de groupes taxonomiques larges et complexes constitués de taxons proches comme chez *Rosa*. L'exploitation de génomes complets disponibles publiquement a permis d'extraire 1856 courtes séquences orthologues en simple copie (SCO_{Tag}s) d'intérêt phylogénétique. Quatre-vingt-douze SCO_{Tag}s du génome nucléaire et quatre SCO_{Tag}s du génome chloroplastique ont été ciblé chez 126 espèces en utilisant des PCR mircrofluidiques et du séquençage d'amplicons. La quantité importante de données générées par le séquençage a permis d'estimer le niveau de ploïdie de chaque accession et d'assembler des séquences alléliques qui ont plus tard servi à tracer l'origine hybride de certains taxons. Une approche par étapes a été développée pour progressivement dévoiler les patterns réticulés chez *Rosa*. Des phylogénies nucléaires et chloroplastiques robustes ont été obtenues ainsi que des scenarios d'hybridation détaillés pour plusieurs spécimens. Enfin, le pouvoir de résolution de marqueurs microsatellite a été étudié pour délimiter des espèces très proches. De nombreux groupes taxonomiques larges et complexes peuvent désormais être étudiés en utilisant cette approche progressive.

Title : Phylogenomics of the genus *Rosa*: Hybridization and polyploidy as factors for diversification

Mots clés : Phylogénomique; *Rosa* sp.; amplicon sequencing; orthology assignment; networks

Abstract : The genus *Rosa* comprises 150-200 species well-distributed throughout the northern hemisphere and presents a complex evolutionary history. Both hybridization and polyploidy represent major driving forces in *Rosa*, yet these two processes were barely investigated in previous phylogenetic studies. With the recent acquisition of whole genome sequencing data and the development of high-throughput sequencing (HTS) techniques, the objective was to develop a general phylogenomics framework to address the phylogenetic relationships in large taxonomic groups made of closely related taxa such in Rosa. A mining strategy first identified 1856 informative single-copy orthologous tags (SCO_{Tag}s) in publicly available whole genome sequencing datasets. Ninety-two SCO_{Tag}s from the nuclear genome and four SCO_{Tag}s from the chloroplast genome were then targeted using microfluidic PCRs and amplicon sequencing in a broader sampling of *Rosa* representing 126 species. The HTS data obtained for each accession enabled to estimate ploidy levels and to assemble allelic sequences that further served to trace the origin of hybrid taxa. A stepwise strategy was developed to gradually unveil the reticulate patterns in *Rosa*. Robust plastid and nuclear phylogenies were obtained as well as detailed hybridization scenarios for several specimens. Finally, the resolving power of microsatellite markers was investigated to delineate close species relationships. Using this stepwise framework, many phylogenetic relationships in large and complex taxonomic groups could now be addressed.