



HAL
open science

Mining high-throughput genomic datasets to investigate the evolutionary history of Oleaceae

Pauline Raimondeau

► **To cite this version:**

Pauline Raimondeau. Mining high-throughput genomic datasets to investigate the evolutionary history of Oleaceae. Biodiversity and Ecology. Université Paul Sabatier - Toulouse III, 2022. English. NNT : 2022TOU30290 . tel-04386753

HAL Id: tel-04386753

<https://theses.hal.science/tel-04386753>

Submitted on 11 Jan 2024

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

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



THÈSE

**En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE
Délivré par l'Université Toulouse 3 - Paul Sabatier**

**Présentée et soutenue par
Pauline RAIMONDEAU**

Le 2 décembre 2022

**Exploration de données génomiques haut-débit pour l'étude de
l'histoire évolutive des Oléacées**

Ecole doctorale : **SEVAB - Sciences Ecologiques, Vétérinaires, Agronomiques et
Bioingenieries**

Spécialité : **Ecologie, biodiversité et évolution**

Unité de recherche :
EDB - Evolution et Diversité Biologique

Thèse dirigée par
Guillaume BESNARD

Jury

Mme Tanja SLOTTE, Rapporteur
M. Benoit NABHOLZ, Rapporteur
M. Guillaume BESNARD, Directeur de thèse
M. CHRISTOPHE THEBAUD, Président

Remerciements

En premier lieu, je remercie Guillaume Besnard qui m'a offert les moyens de faire cette thèse. Merci également à France Olive et à la Région Occitanie d'avoir financé ce projet.

Un très grand merci à Tanja Slotte, Benoit Nabholz et Christophe Thebaud d'avoir accepté d'examiner mon travail.

J'exprime ma plus profonde gratitude envers Pascal-Antoine Christin qui est pour moi un exemple de par ses qualités scientifiques et humaines. Je ne serais pas aujourd'hui « docteur » sans cette rencontre. Merci pour l'aide si précieuse. Merci pour la positivité !

Je tiens à remercier Pierre-Marc Delaux de m'avoir hébergé dans sa belle équipe EVO. Merci de m'avoir redonné confiance.

Merci en particulier au bureau bioinfo pour l'ambiance inoubliable. Merci de m'avoir fait une place, pris sous votre aile tels de nobles kakapos (eh oui, je l'ai mis). En quelques jours, j'étais chez moi. Merci pour l'aide et les encouragements. Merci pour les rires. Merci Camille, Chloé, Cyril, Jean et Maxime. Vous méritez plus de gâteaux!

Merci à mes amis "du labo" Alex, Déborah, Julia, Louise, Opale, Ricardo, Sean et Thomas. Quelle chance d'avoir fait ce chemin ensemble ! Merci d'avoir égayé ces trois années.

Merci à Candice et à Constance pour le soutien indéfectible. Les moments passés ensemble ont été des rayons de soleil et pas seulement lors de vos visites dans le midi toulousain !

Merci à ma famille et tout particulièrement ma sœur Laura pour l'écoute et la patience. Je sais que ça a été long.

Cette thèse, je la dois à toutes ces personnes formidables qui m'ont épaulée. C'est la découverte la plus importante que j'ai faite !

Table of contents

Introduction	1
Chapter 1. Resolving the phylogeny of the olive family: confronting information from organellar and nuclear genomes	13
Appendix for Chapter 1	32
Chapter 2. Positive selection on non-photosynthetic genes drives parallel acceleration of plastid genome evolution in two Oleaceae lineages	41
Appendix for Chapter 2	64
Chapter 3. A hemizygous supergene controls homomorphic di-allelic self-incompatibility in olive	74
Appendix for Chapter 3	91
Chapter 4. Transcriptomics of the development of distyly in jasmine	99
Appendix for Chapter 4	111
Chapter 5. Tracking the spatio-temporal spread of the cultivated olive with archaeogenomics	118
Appendix for Chapter 5	139
Conclusion and perspectives	147

Introduction

I. The genomics era

The genesis of DNA sequencing

In 1869, a mysterious phosphate-rich substance was isolated in cell nuclei. Fifty years later, it was proposed that this “nucleic acid” was made of four phosphate bases and sugar (“nucleotides”) linked together through phosphate groups. The demonstration that this molecule was the constituent of genes and thus, hereditary information came twenty-five years later. A fantastic accumulation of both theoretical and experimental works finally led to the emergence of the central dogma of biology (Crick, 1958). The whole framework revolves around the transfer of one key element, the sequence information: DNA carries genetic information encoded in the form of a precisely ordered succession of nucleotides. The sequence is the information. Moving from sequence composition to precise order was challenging. The first whole nucleic acid sequence, a tRNA consisting of 16 nucleotides was eventually obtained in, 1965 (Holley *et al.*, 1965). Numerous other methods were developed in parallel to sequence DNA, but all were extremely labor-intensive and restricted to short stretches of nucleic acid, until the development of Sanger's ‘chain-termination’ in 1977 (Sanger *et al.*, 1977). The first complete genomes were published a few years later and the accuracy, robustness and ease of use made Sanger sequencing the standard sequencing method for decades.

Genetics to a new scale

DNA sequencing transformed biological sciences and has applications in a vast range of fields such as medical diagnosis and treatments, forensic, crop breeding or systematics. Access to genetic information more broadly accelerated our understanding of organisms diversity and functions. Nevertheless, as Sanger sequencing can produce reads of around 1 kb at most, sequencing longer DNA fragments requires cloning overlapping DNA fragments prior to sequencing. Despite automation and technical improvement, sequencing large genomes with this technique remains prohibitively expensive and laborious. This limitation called for the development of high-throughput sequencing methods, today known as second and third sequencing generations. Most of these methods follow the same principle as Sanger sequencing (sequencing-by-synthesis). The major breakthrough of these high-throughput techniques is the parallelisation of thousands of synthesis reactions thanks to miniaturisation (Mardis, 2017). Their introduction 15 years ago has led to a major shift in biology from genetics to genomics. As measured as sequencing throughput increased, sequencing cost plummeted, offering the possibility to sequence complete genomes for a

vast diversity of organisms (Figure 1). Our understanding of phenotypes is no longer limited by the availability of few targeted genes sequences. It is now possible to investigate multi-factorial traits using thousands markers in a large number of individuals. DNA sequencing has become a key technology in numerous fields and gave birth to genomics (Goodwin et al., 2016). The progress in sequencing capacity has been exponential, and as led to an explosion of genetic data availability (Figure 2). Since 1982, the number of DNA bases in public repositories has doubled approximately every 18 months (NCBI).

The accumulation of genomics data has led some authors to use the term “data deluge” and claim that our ability to sequence DNA is outpacing our ability to decipher the information it contains (Schatz and Langmead, 2013). Indeed, the most costly part has moved from generating the genetic data to storing and analyzing them (Sboner *et al.*, 2011). This sudden transition to genome-scale data has prompted to major innovations in computational methods and bioinformatics. The development of new approaches to analyse large biological datasets is one of the most vibrant fields of research today.

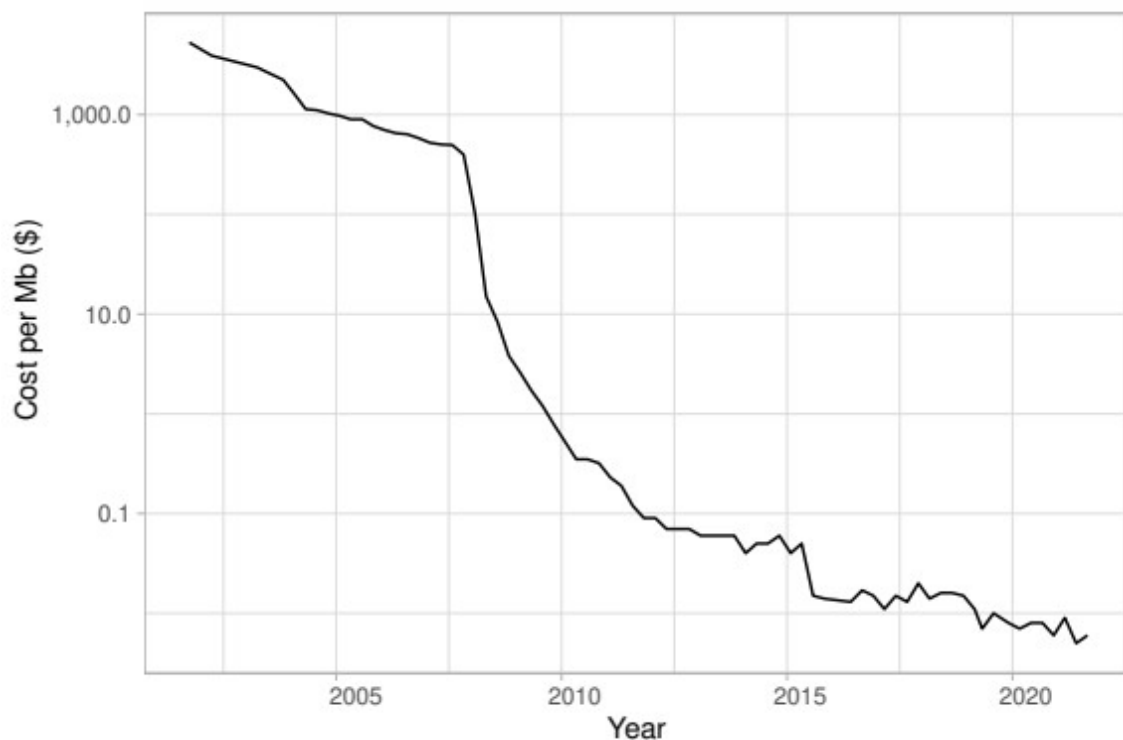


Figure 1. Cost for sequencing one megabase of DNA over year. Data from the National Human Genome Research Institute (Wetterstrand, 2022).

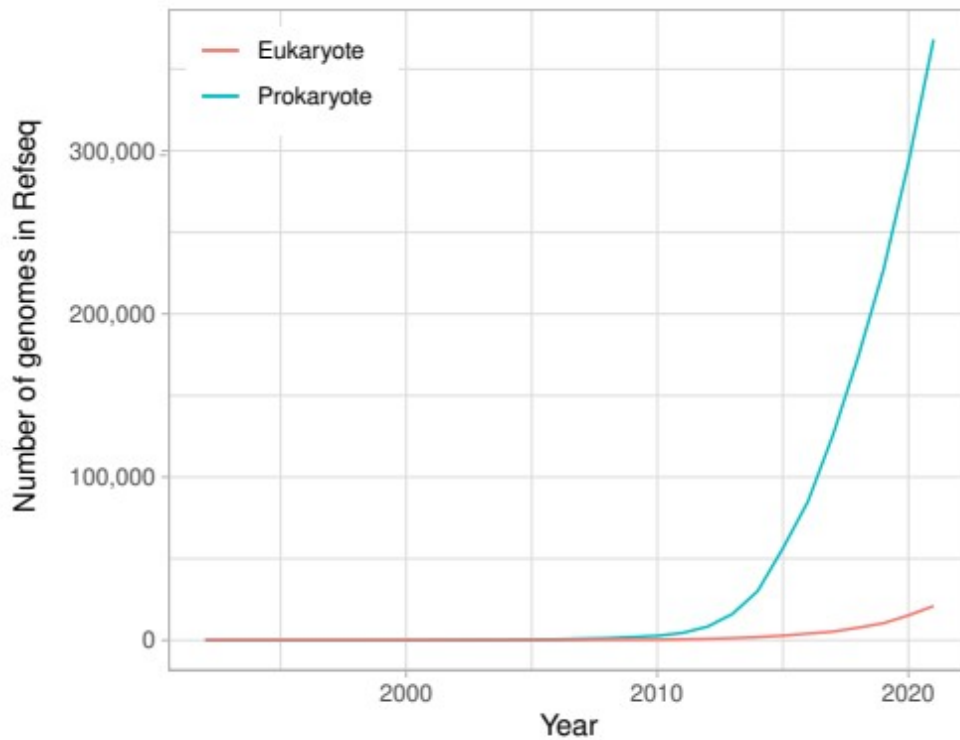


Figure 2. Accumulation of published genome for prokaryotes and eukaryotes in RefSeq. Genome release dates were extracted from RefSeq assembly summary file (August 2022 release).

Genomics contribution to study of evolution

The idea that species were not fixed but can evolve over time slowly emerged at the beginning of the 19th century with Lamarck's transformism. Darwin's theory of descent with modification in 1859 proposed the idea of a branching tree of life, where different species share a common ancestor. The amount of evidence he gathered progressively led to the acceptance of the general concept of evolution but not that of natural selection as the mechanism of evolution. Only the integration of genetics enabled the establishment of a widely accepted theory of evolution in the 1930s and 1940s. Evolutionary biology was greatly transformed by the rise of molecular genetics. The subfields of molecular evolution rapidly revealed the importance of neutral change and drift and molecular phylogenetics reorganized the tree of life. The resolving power was, however, rapidly limited by the availability of genetic data (few genes/few species/few individuals). Again, genomics offered a new scale at which to study the change of living organisms across time and the relationships between them. The rise of high-throughput sequencing empowered the discipline and comparative genomics now has a central position in biology. Phylogenetics and molecular evolution is not confined to evolutionary biology anymore but suffuse numerous fields in biology either functional studies, ecology, biogeography, taxonomy, paleontology or conservation. New questions are now

addressable and evolutionary biologist try to understand the timing, scale, effects and interactions of genomic changes in evolutionary processes such as speciation, adaptation or domestication.

II. Study system: the Oleaceae family

This thesis focused on the evolution of Oleaceae, a particularly interesting family of plants where several of these questions remain completely unexplored with genomics data. Oleaceae is a medium-sized family encompassing about 700 species widely distributed in temperate or tropical regions of every continent (Green, 2004). It notably includes some genus of economic or ornamental value such as olive, ash, jasmine, privet, forsythia or lilac.

A diversity of life-history traits

The wide distribution of Oleaceae species in diverse habitats only mirrors the great diversity of morphological and life-history traits in the family. First, Oleaceae can take a variety of life forms: shrubs, trees, a few lianas and even one herbaceous species have been described (Green, 2002). Oleaceae also exhibits an impressive diversity of floral traits (Sehr and Weber, 2009) and fruit types (e.g. capsules, samaras, drupes, berries; Figure 3; Rohwer, 1993), probably influenced by ecological shifts (Hinsinger *et al.*, 2013). Finally, an interesting assembly of mating systems has been described in the family, from hermaphroditism to dioecy, with several intermediate stages such as polygamy and androdioecy (Wallander, 2008; Saumitou-Laprade *et al.*, 2018). This variety of sex systems comes together with different breeding strategies. Most Oleaceae species are outcrossing but some are capable of self-fertilization (Lepart and Dommée, 1992; Dommée *et al.*, 1999). The variety of mating system in Oleaceae has been linked to the presence of a peculiar self-incompatibility system (Saumitou-Laprade *et al.*, 2010) with only two homomorphic compatibility groups. This peculiar system is thought to have enabled the stable maintenance of androdioecy in *Phillyrea* but also to have facilitated transitions to dioecy in some *Fraxinus* (Vernet *et al.*, 2016). This is to date the only reported occurrence of a homomorphic di-allelic self-incompatibility system in plants. Heteromorphic self-incompatibility systems (approach herkogamy, distyly) are also described in the family (Thompson and Dommée, 2000; Hong and Han, 2002; Olesen *et al.*, 2003; Ganguly and Barua, 2020).



Figure 3. Diversity of floral (top row) and fruit (bottom row) type among Oleaceae species. From left to right: *Olea europaea*, *Fraxinus excelsior*, *Crysojasmium fruticans*, *Norhonia emarginata*. Photo credit (from top left to bottom right): P. Dietze, B. Kepic, T. Chugg, N. Juillet, RBG Kew, Y. Le Bastard, C. Perez, K. de Niort.

Evolutionary genomics of Oleaceae

The family experienced two recent whole-genome duplication events: a first one common with its sister family Carlemanniaceae (with which Oleaceae form a lineage sister to all other Lamiales), and a second one within the family, at the base of the tribe Oleae (Zhang *et al.*, 2020). This state of paleopolyploids reflects on species chromosome number which is 23 in Oleae and ranges from 11 to 14 in other Oleaceae (Taylor, 1945). Most work on Oleaceae's systematics preceded the genomic revolution and were based on morphological, cytological, and biochemical traits (Taylor, 1945; Harborne and Green, 1980). Despite the increasing amount of genomic-scale datasets generated for Oleaceae species, the family evolutionary history has remained largely unexplored with these methods.

The olive tree (*Olea europaea*) is probably the most studied Oleaceae: the first plastome was published in 2011 (Mariotti *et al.*, 2010) and the first nuclear genome in 2016 (Cruz *et al.*, 2016). A first genome for ash (*Fraxinus excelsior*) was also published that year (Sollars *et al.*, 2017). Sequencing efforts for Oleaceae species accelerated in the past few years and nuclear genomes are now available for *Osmanthus fragrans* (Yang *et al.*, 2018), *Forsythia suspensa* (Li *et al.*, 2020), *Jasminum sambac* (Chen *et al.*, 2020) and 22 other *Fraxinus* species (Stocks *et al.*, 2019). These genomes are of various qualities but reflects the accumulation of genomics data for this family. More than a hundred plastomes are now publicly available as well as a few mitochondrial genomes. The data are now available to enter Oleaceae evolution study in the genomic era.

Phylogenomics will enable revisions of the evolutionary relationships in the family. Indeed, molecular phylogenies for Oleaceae produced during the last two decades were mainly focused on specific lineages (Lee *et al.*, 2007; Yuan *et al.*, 2010; Kim and Kim, 2011; Hong-Wa and Besnard, 2013; Ha *et al.*, 2018; Olofsson *et al.*, 2019). The last family wide phylogenetic work was that of Wallander & Albert (2000), based on two plastid non-coding markers. It defined 5 tribes: Myxopyreae, Fontanesieae, Forsythieae, Jasmineae and Oleeeae (Figure 4), as well as a division in 4 subtribes within Oleeeae (Ligustrinae, Schreberinae, Fraxininae and Oleinae).

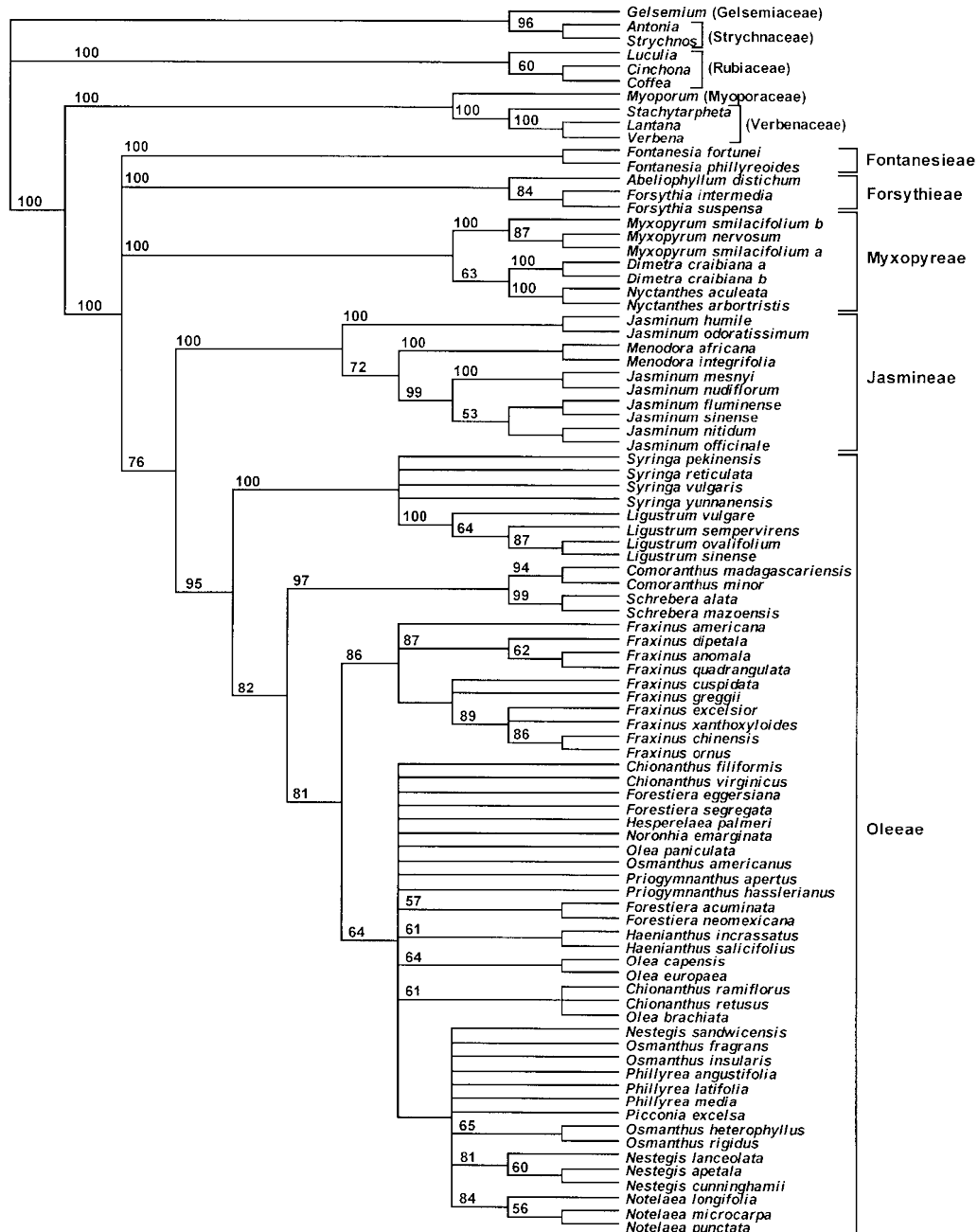


Figure 4. Phylogenetic tree of Oleaceae (from Wallander and Albert, 2000). Strict consensus of the most parsimonious trees with jackknife support values (only >50% are shown).

The relationship at lower levels were not well resolved and the sampling not exhaustive, mostly due to the difficulty to sample rare tropical species. Development of museomics enables the sequencing of inaccessible specimens and the inclusion in genomic-scale datasets to reach a better resolution of the evolutionary relationship between lineages, as exemplified for a grass lineage in Silva *et al.* (2017).

Moreover, the analysis of ancient DNA (aDNA) extracted from archaeological samples can offer unprecedented possibility to trace olive domestication process. The evolutionary history of olive has been studied at length (see Besnard *et al.*, 2018 for review). Yet, some of its aspects are still not understood. In particular, the number of independent domestication events is debated (Diez *et al.*, 2015; Besnard and Rubio de Casas, 2016). Molecular markers have been used to study the distribution of distinct lineages and infer their trajectories but the reliance on extant olive diversity is a limiting factor to arbitrate definitely between the different suggested scenario. Archaeological samples can provide additional anchor points to track olive domestication in space and time.

Phylogenomics is also a great tool to study variation in evolutionary rates. Life-history traits can have a profound impact on rates of molecular evolution (Nabholz *et al.*, 2008; Smith and Donoghue, 2008; Thomas *et al.*, 2010; Lanfear *et al.*, 2013; Bromham *et al.*, 2015). The diversity of traits described in the previous section offers a great system to test their potential impact on evolutionary rates.

Comparative genomics is also a powerful tool to tackle the most intriguing question that emerged in the family at the discovery of the homomorphic di-allelic self-incompatibility (Saumitou-Laprade *et al.*, 2010; Vernet *et al.*, 2016). Most of what we know about this system comes from experimental studies of crossing (Saumitou-Laprade *et al.*, 2010; Vernet *et al.*, 2016; Saumitou-Laprade *et al.*, 2017; Besnard *et al.*, 2020; De Cauwer *et al.*, 2020) and only two papers, published in the past two years, investigate the genetics underlying incompatibility phenotypes (Mariotti *et al.*, 2020; Carré *et al.*, 2021). Still very little is known on the genetic determinism, as well as regarding the evolution and maintenance of this system. Findings can have important implications both fundamentally, on our understanding of plant mating-system evolution, and economically, for olive production.

III- Thesis outline

As part of this work, we leverage the increasing availability of genomics information to investigate several pending questions about the evolution of Oleaceae.

In the first chapter, we first provide a revision of the family phylogeny, an important prerequisite to the study of trait diversification.

In the second chapter, we investigate variation in molecular evolutionary rates between lineages and genomic compartments (nuclear, chloroplastic or mitochondrial). The phylogeny obtained in the first chapter indeed revealed important variations in branch lengths in some lineages and we thus undertook a precise evaluation of this phenomenon. Our findings posit that a slight change during reproduction can impact spectacularly the evolutionary trajectory of plastid genomes.

In the third chapter, we dive into the fascinating homomorphic di-allelic self-incompatibility system present in olive. We combine genomics evidence for three subspecies of olive to uncover the genetic determinism of self-incompatibility phenotype.

The results of a transcriptomics study of jasmine flowers are presented in a fourth chapter. The aim of this part was to identify transcripts involved in the distylous self-incompatibility system and potentially inform us on the homology of this di-allelic system to the one studied in the previous chapter.

Finally, the last chapter consists of a pilot study for the use of archeogenetics to study olive domestication. We show that ancient DNA recovered from archeological samples from around the Mediterranean basin can bring new insights on the domestication history of olive.

References

- Besnard G, Cheptou PO, Debbaoui M, Lafont P, Hugueny B, Dupin J, Baali-Cherif D. 2020. Paternity tests support a diallelic self-incompatibility system in a wild olive (*Olea europaea* subsp. *laperrinei*, *Oleaceae*). *Ecol. Evol.*:1876–1888.
- Besnard G, Terral JF, Cornille A. 2018. On the origins and domestication of the olive: A review and perspectives. *Ann. Bot.* 121:385–403.
- Besnard G, Rubio de Casas R. 2016. Single vs multiple independent olive domestications: The jury is (still) out. *New Phytol.* 209:466–470.
- Bromham L, Hua X, Lanfear R, Cowman PF. 2015. Exploring the relationships between mutation rates, life history, genome size, environment, and species richness in flowering plants. *Am. Nat.* 185:508–524.
- Carré A, Gallina S, Santoni S, Vernet P, Godé C, Castric V, Saumitou-Laprade P. 2021. Genetic mapping of sex and self-incompatibility determinants in the androdioecious plant *Phillyrea angustifolia*. *Peer Community J.* 1:1–20.
- De Cauwer I, Vernet P, Billiard S, Godé C, Bourceaux A, Ponitzki C, Saumitou-Laprade P. 2020. Widespread coexistence of self-compatible and self-incompatible phenotypes in a diallelic self-incompatibility system in *Ligustrum vulgare* (*Oleaceae*).
- Chen G, Mostafa S, Lu Z, Du R, Cui J, Wang Y, Liao Q, Lu J, Mao X, Chang B, *et al.* 2020. The jasmine (*Jasminum sambac*) genome and flower fragrances. *bioRxiv*:0–2.
- Crick FH. 1958. On protein synthesis. *Symp. Soc. Exp. Biol.* 12:138–163.
- Cruz F, Julca I, Gómez-Garrido J, Loska D, Marcet-Houben M, Cano E, Galán B, Frias L, Ribeca P, Derdak S, *et al.* 2016. Genome sequence of the olive tree, *Olea europaea*. *Gigascience* 5:29.
- Diez CM, Trujillo I, Martínez-Urdiroz N, Barranco D, Rallo L, Marfil P, Gaut BS. 2015. Olive domestication and diversification in the Mediterranean Basin. *New Phytol.* 206:436–447.
- Dommée B, Geslot A, Thompson JD, Reille M, Denelle N. 1999. Androdioecy in the entomophilous tree *Fraxinus ornus* (*Oleaceae*). *New Phytol.* 143:419–426.
- Ganguly S, Barua D. 2020. High herkogamy but low reciprocity characterizes isoplethic populations of *Jasminum malabaricum*, a species with stigma-height dimorphism. *Plant Biol.* 22:899–909.
- Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17:333–351.
- Green PS. 2002. A Revision of *Olea* L. (*Oleaceae*). *Kew Bull.* 57:91–140.
- Green, P.S. *Oleaceae*. In *The Families and Genera of Vascular Plants, Flowering Plants, Dicotyledons*; Kubitzki, K., Kadereit, J.W., Eds.; Springer: New York, NY, USA, 2004; Volume 7, pp. 296–306.
- Ha Y-H, Kim C, Choi K, Kim J-H. 2018. Molecular Phylogeny and Dating of *Forsythieae* (*Oleaceae*) Provide Insight into the Miocene History of Eurasian Temperate Shrubs. *Front. Plant Sci.* 9:99.
- Harborne JB, Green PS. 1980. A chemotaxonomic survey of flavonoids in leaves of the *Oleaceae*. *Bot. J. Linn. Soc.* 81:155–167.

- Hinsinger DD, Basak J, Gaudeul M, Cruaud C, Bertolino P, Frascaria-Lacoste N, Bousquet J. 2013. The Phylogeny and Biogeographic History of Ashes (*Fraxinus*, Oleaceae) Highlight the Roles of Migration and Vicariance in the Diversification of Temperate Trees. *PLoS One*8:e80431.
- Holley RW, Apgar J, Everett GA, Madison JT, Marquisee M, Merrill SH, Penswick JR, Zamir A. 1965. Structure of a ribonucleic acid. *Science* (80-). 147:1462–1465.
- Hong-Wa C, Besnard G. 2013. Intricate patterns of phylogenetic relationships in the olive family as inferred from multi-locus plastid and nuclear DNA sequence analyses: A close-up on *Chionanthus* and *Noronhia* (Oleaceae). *Mol. Phylogenet. Evol.*67:367–378.
- Hong SP, Han MJ. 2002. The floral dimorphism in the rare endemic plant, *Abeliophyllum distichum* Nakai (Oleaceae). *Flora* 197:317–325.
- Kim D-K, Kim J-H. 2011. Molecular phylogeny of tribe Forsythieae (Oleaceae) based on nuclear ribosomal DNA internal transcribed spacers and plastid DNA trnL-F and matK gene sequences. *J. Plant Res.*124:339–347.
- Lanfear R, Ho SYW, Jonathan Davies T, Moles AT, Aarssen L, Swenson NG, Warman L, Zanne AE, Allen AP. 2013. Taller plants have lower rates of molecular evolution. *Nat. Commun.* 4:1–7.
- Lee H-L, Jansen RK, Chumley TW, Kim K-J. 2007. Gene Relocations within Chloroplast Genomes of *Jasminum* and *Menodora* (Oleaceae) Are Due to Multiple, Overlapping Inversions. *Mol. Biol. Evol.*24:1161–1180.
- Lepart J, Dommée B. 1992. Is *Phillyrea angustifolia* L. (Oleaceae) an androdioecious species? *Bot. J. Linn. Soc.*108:375–387.
- Li L-F, Cushman SA, He Y-X, Li Y. 2020. Genome sequencing and population genomics modeling provide insights into the local adaptation of weeping forsythia. *Hortic. Res.*7:130.
- Mardis ER. 2017. DNA sequencing technologies: 2006-2016. *Nat. Protoc.* 12:213–218.
- Mariotti R, Cultrera NGM, Diez CM, Baldoni L, Rubini A. 2010. Identification of new polymorphic regions and differentiation of cultivated olives (*Olea europaea* L.) through plastome sequence comparison. *BMC Plant Biol.*10:211.
- Mariotti R, Fornasiero A, Mousavi S, Cultrera NGM, Brizioli F, Pandolfi S, Passeri V, Rossi M, Magris G, Scalabrin S, *et al.* 2020. Genetic Mapping of the Incompatibility Locus in Olive and Development of a Linked Sequence-Tagged Site Marker. *Front. Plant Sci.* 10:1–13.
- Nabholz B, Glémin S, Galtier N. 2008. Strong variations of mitochondrial mutation rate across mammals - The longevity hypothesis. *Mol. Biol. Evol.* 25:120–130.
- NCBI. Genbank and Whole Genome Sequencing statistics. Available at www.ncbi.nlm.nih.gov/genbank/statistics/. Accessed: August 2022.
- Olesen JM, Dupont YL, Ehlers BK, Valido A, Hansen DM. 2003. Heterostyly in the Canarian endemic *Jasminum odoratissimum* (Oleaceae). *Nord. J. Bot.* 23:537–539.
- Olofsson JK, Cantera I, Van de Paer C, Hong-Wa C, Zedane L, Dunning LT, Alberti A, Christin PA, Besnard G. 2019. Phylogenomics using low-depth whole genome sequencing: A case study with the olive tribe. *Mol. Ecol. Resour.* 19:877–892.
- Rohwer, J.G. A preliminary survey of the fruits and seeds of the Oleaceae. *Bot. Jahrb. Syst.* 1993, 115, 271–291.
- Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* 74:5463–5467.

- Saumitou-Laprade P, Vernet P, Dowkiw A, Bertrand S, Billiard S, Albert B, Gouyon P-H, Dufay M. 2018. Polygamy or subdioecy? The impact of diallelic self-incompatibility on the sexual system in *Fraxinus excelsior* (Oleaceae). *Proc. R. Soc. B Biol. Sci.* 285:20180004.
- Saumitou-Laprade P, Vernet P, Vassiliadis C, Hoareau Y, De Magny G, Dommée B, Lepart J. 2010. A self-incompatibility system explains high male frequencies in an androdioecious plant. *Science (80-.)*. 327:1648–1650.
- Saumitou-Laprade P, Vernet P, Vekemans X, Billiard S, Gallina S, Essalouh L, Mhaïa A, Moukhli A, El Bakkali A, Barcaccia G, *et al.* 2017. Elucidation of the genetic architecture of self-incompatibility in olive: Evolutionary consequences and perspectives for orchard management. *Evol. Appl.* 10:867–880.
- Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. 2011. The real cost of sequencing: higher than you think! *Genome Biol.* 12:125.
- Schatz MC, Langmead B. 2013. The DNA Data Deluge: Fast, efficient genome sequencing machines are spewing out more data than geneticists can analyze. *IEEE Spectr* 50:26—33.
- Sehr EM, Weber A. 2009. Floral ontogeny of oleaceae and its systematic implications. *Int. J. Plant Sci.* 170:845–859.
- Silva C, Besnard G, Piot A, Razanatsoa J, Oliveira RP, Vorontsova MS. 2017. Museomics resolve the systematics of an endangered grass lineage endemic to north-western Madagascar. *Ann. Bot.* 119:339–351.
- Smith SA, Donoghue MJ. 2008. Rates of molecular evolution are linked to life history in flowering plants. *Science (80-.)*. 322:86–89.
- Sollars ESA, Harper AL, Kelly LJ, Sambles CM, Ramirez-Gonzalez RH, Swarbreck D, Kaithakottil G, Cooper ED, Uauy C, Havlickova L, *et al.* 2017. Genome sequence and genetic diversity of European ash trees. *Nature* 541:212–216.
- Stocks JJ, Metheringham CL, Plumb WJ, Lee SJ, Kelly LJ, Nichols RA, Buggs RJA. 2019. Genomic basis of European ash tree resistance to ash dieback fungus. *Nat. Ecol. Evol.* 3:1686–1696.
- Taylor H. 1945. Cyto-taxonomy and phylogeny of the Oleaceae. *Brittonia* 5:337–367.
- Thomas JA, Welch JJ, Lanfear R, Bromham L. 2010. A generation time effect on the rate of molecular evolution in invertebrates. *Mol. Biol. Evol.* 27:1173–1180.
- Thompson JD, Dommée B. 2000. Morph-specific patterns of variation in stigma height in natural populations of distylous *Jasminum fruticans*. *New Phytol.* 148:303–314.
- Vernet P, Lepercq P, Billiard S, Bourceaux A, Lepart J, Dommée B, Saumitou-Laprade P. 2016. Evidence for the long-term maintenance of a rare self-incompatibility system in Oleaceae. *New Phytol.* 210:1408–1417.
- Wallander E. 2008. Systematics of *Fraxinus* (Oleaceae) and evolution of dioecy. *Plant Syst. Evol.* 273:25–49.
- Wallander E, Albert VA. 2000. Phylogeny and classification of Oleaceae based on *rps16* and *trnL-F* sequence data. *Am. J. Bot.* 87:1827–1841.
- Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. Accessed: August 2022.

- Yang X, Yue Y, Li H, Ding W, Chen G, Shi T, Chen J, Park MS, Chen F, Wang L. 2018. The chromosome-level quality genome provides insights into the evolution of the biosynthesis genes for aroma compounds of *Osmanthus fragrans*. *Hortic. Res.*5:72.
- Yuan W-J, Zhang W-R, Han Y-J, Dong M-F, Shang F-D. 2010. Molecular phylogeny of *Osmanthus* (Oleaceae) based on non-coding chloroplast and nuclear ribosomal internal transcribed spacer regions. *J. Syst. Evol.*48:482–489.
- Zhang C, Zhang T, Luebert F, Xiang Y, Huang C-H, Hu Y, Rees M, Frohlich MW, Qi J, Weigend M, *et al.* 2020. Asterid Phylogenomics/Phylotranscriptomics Uncover Morphological Evolutionary Histories and Support Phylogenetic Placement for Numerous Whole-Genome Duplications. *Mol. Biol. Evol.*37:3188–3210.

Chapter 1

Resolving the phylogeny of the olive family (Oleaceae): confronting information from organellar and nuclear genomes



Dupin J.*, Raimondeau P.*, Hong-Wa C., Manzi S., Gaudeul M., Besnard G.

*These authors contributed equally

Genes 2020, 11, 1508

Article

Resolving the Phylogeny of the Olive Family (Oleaceae): Confronting Information from Organellar and Nuclear Genomes

Julia Dupin ^{1,†}, Pauline Raimondeau ^{1,†}, Cynthia Hong-Wa ², Sophie Manzi ¹ ,
Myriam Gaudeul ³ and Guillaume Besnard ^{1,*} 

¹ Laboratoire Evolution & Diversité Biologique (EDB, UMR 5174), CNRS/IRD/Université Toulouse III, 118 Route de Narbonne, 31062 Toulouse, France; julia.guedes-rocha-dupin@univ-tlse3.fr (J.D.); pauline.raimondeau@univ-tlse3.fr (P.R.); sophie.manzi@univ-tlse3.fr (S.M.)

² Claude E. Phillips Herbarium, Delaware State University, 1200 N. Dupont Hwy, Dover, DE 19901-2277, USA; chwa@desu.edu

³ Institut de Systématique Evolution Biodiversité (ISYEB), Muséum National d'Histoire Naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles, 57 rue Cuvier, CP39, 75005 Paris, France; myriam.gaudeul@mnhn.fr

* Correspondence: guillaume.besnard@univ-tlse3.fr

† These authors contributed equally to this work.

Received: 25 September 2020; Accepted: 11 December 2020; Published: 16 December 2020



Abstract: The olive family, Oleaceae, is a group of woody plants comprising 28 genera and ca. 700 species, distributed on all continents (except Antarctica) in both temperate and tropical environments. It includes several genera of major economic and ecological importance such as olives, ash trees, jasmines, forsythias, osmanthus, privets and lilacs. The natural history of the group is not completely understood yet, but its diversification seems to be associated with polyploidisation events and the evolution of various reproductive and dispersal strategies. In addition, some taxonomical issues still need to be resolved, particularly in the paleopolyploid tribe Oleae. Reconstructing a robust phylogenetic hypothesis is thus an important step toward a better comprehension of Oleaceae's diversity. Here, we reconstructed phylogenies of the olive family using 80 plastid coding sequences, 37 mitochondrial genes, the complete nuclear ribosomal cluster and a small multigene family encoding phytochromes (*phyB* and *phyE*) of 61 representative species. Tribes and subtribes were strongly supported by all phylogenetic reconstructions, while a few Oleae genera are still polyphyletic (*Chionanthus*, *Olea*, *Osmanthus*, *Nestegis*) or paraphyletic (*Schrebera*, *Syringa*). Some phylogenetic relationships among tribes remain poorly resolved with conflicts between topologies reconstructed from different genomic regions. The use of nuclear data remains an important challenge especially in a group with ploidy changes (both paleo- and neo-polyploids). This work provides new genomic datasets that will assist the study of the biogeography and taxonomy of the whole Oleaceae.

Keywords: herbarium; museum collection; mitochondrial DNA; plastome; nuclear ribosomal DNA; phytochromes; low-copy genes; taxonomy; polyploidy

1. Introduction

The olive family (Oleaceae) is a medium-sized group of woody plants comprising 28 genera and ca. 700 species, distributed on all continents (except Antarctica) in both temperate and tropical environments [1]. Most species are trees, but there are also one herbaceous plant (*Dimetra craibiana*), small shrubs (e.g., *Menodora* spp.) and a few lianas (e.g., *Jasminum* spp., *Chionanthus macrobotrys*).

Many Oleaceae species are of economic importance, for the production of oil and fruits (olive), timber (e.g., ash trees), as well as ornaments and fragrances (e.g., jasmines, osmanthus, lilacs, etc). Moreover, Oleaceae are important components of temperate and tropical ecosystems, with, for example, several species producing drupes and palatable leaves as a common food source to wild animals. Individual species can also support a large number of other organisms, for example, nearly 1000 species (e.g., fungi, insects, birds) are known to be associated with *Fraxinus excelsior* [2].

Oleaceae is currently divided into five tribes, Myxopyreae, Jasmineae, Forsythieae, Fontanesieae and Oleae, the latter being subdivided into four subtribes (Oleinae, Fraxininae, Ligustrinae, and Schreberinae) [3]. The natural history of the group is not completely understood yet, but its diversification seems to be associated with a few events of polyploidisation (in particular a major event of whole genome duplication in the ancestor of Oleae) [4–9] and the evolution of various reproductive and dispersal strategies [3]. This family thus presents a substantial diversity of flowers (e.g., [1,10,11]) and fruits (e.g., capsules, samaras, drupes) [3,12–14]; associated with different vectors of pollination and seed dispersal. Variable breeding systems were also described, from hermaphroditism to dioecy, with several stages often considered as intermediate such as polygamy and androdioecy (e.g., [10,15–18]). In addition, a di-allelic self-incompatibility system, associated with distyly in some groups, has been reported for a number of species belonging to different tribes (i.e., Myxopyreae [15], Jasmineae [19,20], Forsythieae [21], and Oleae [1,22,23]).

To better understand trait evolution and patterns of diversification in this group, or to resolve any lingering taxonomical issues, as in the case of the paleopolyploid tribe Oleae [3,24–26], a robust phylogenetic hypothesis is then required. Oleaceae's systematics has evolved from being based on morphological, cytological, and biochemical traits (e.g., [4,14,27]), to the use of molecular phylogenies during the last two decades. Such advances, though, have mainly consisted in studies that focused on specific groups or partially resolved phylogenetic trees [11,26,28–36], and not on the whole family (but see [3,37]). This has been mainly due to difficulties to sample all main Oleaceae lineages and to take into account variable ploidy levels [31]. Recent developments in genomics and museomics present new opportunities to tackle such obstacles, though. Several Oleaceae nuclear and cytoplasmic genomes, as well as transcriptomes have been released [31,38–43]. Also herbarium samples, previously deemed unusable, are now accessible allowing for a more comprehensive sampling, and the inclusion of rare, or recently extinct species [31,44–46].

Another product of the recent advances in genomics is the possibility to use various, independent genomic regions to reconstruct phylogenies of plant groups. The genome skimming approach, for instance, has allowed for cost-effective sequencing of high-copy fractions of total genomic DNA, such as organellar genomes, and nuclear ribosomal DNA, but it can also generate data sets for low-copy nuclear genes [47,48]. With such a diversity of genomic datasets, one can compare the phylogenetic hypothesis estimated using plastid, mitochondrial and nuclear data and increment the estimation of species trees from gene trees. Recent studies that did such comparisons include determining the origin of wild octoploid species [49], the placement of the Celastrales-Oxalidales-Malpighiales (COM) clade within Rosidae [50], and the origin and evolution of species in *Ludwigia* sect. *Macrocarpon* of Onagraceae [51].

Here, we reconstructed phylogenies of the olive family using protein-coding sequences for 80 plastid coding sequences, 37 mitochondrial genes, the nuclear ribosomal cluster and one small multigene family encoding phytochromes (*phyB* and *phyE*) of 61 representative species to document any patterns of incongruence between datasets, and discuss these in the context of the evolution of Oleaceae. The use of nuclear data remains, however, an important challenge especially in a group with frequent ploidy changes (both paleo- and neo-polyploids). Due to paleo-events of polyploidisation, the basic chromosome number varies among tribes (i.e., $x = 23$ in Oleae, 14 in Forsythieae, 13 in Fontanesieae, 11 to 13 in Jasmineae, and 11 in Myxopyreae [3,4,6]). As the consequence of whole genome duplication(s), some nuclear genes could be duplicated in polyploid lineages (e.g., Oleae), and their orthology has to be verified before using them for inferring species phylogenies.

In addition, gene duplicates as well as their pseudogenes may inform us on the polyploids ancestors. Reconstructing the phylogeny of multigene families is the first step to identify gene orthologs that could be used for species phylogenetic reconstruction. Here, we chose the closely related phytochrome genes *phyB* and *phyE*, because these low-copy genes have been frequently used for inferring phylogenetic relationships in several plant families (e.g., [52,53]). All these new datasets will not only assist on the study of Oleaceae's taxonomy, but also its biogeography.

2. Materials and Methods

2.1. Taxon Sampling and Sequencing

In this study, we sampled a total of 65 species: 61 belonging to the ingroup (Table S1) and four representing outgroups (Table S2). The ingroup included species representing all currently recognized tribes, subtribes and genera in Oleaceae. For such list, we followed the current checklist of accepted taxa in Oleaceae that has been reviewed by the staff at Royal Botanic Gardens (Kew), as part of the project "World Checklist of Selected Plant Families" [54], and the most recent literature (e.g., [32,55]). The outgroup comprised two species also in the Lamiales order, *Avicennia marina* (Acanthaceae) and *Sesamum indicum* (Pedaliaceae), and two species in the Solanales order, *Capsicum annuum* and *Solanum lycopersicum* (both in Solanaceae).

Whole genome sequences ('genome skims') were obtained for the 61 Oleaceae species. Twenty-two samples were removed from herbarium collections specimens (Table S1). Forty-one accessions were already characterized from previous works [31,42], and we newly analyzed 20 species belonging to Jasmineae (six species, three genera), Myxopyreae (four species, three genera), Forsythieae (*Abeliophyllum*), and Oleae (two accessions of *Ligustrum*, one of *Chengiodendron*, *Chionanthus*, *Haenianthus*, *Syringa*, *Priogymnanthus*, *Noronhia*, and *Comoranthus*). For these samples, total genomic DNA was extracted from ca. 5–10 mg of dried leaves. We grounded the samples in 2-mL tubes with three metal beads using a TissueLyser (Qiagen Inc., Texas). We then extracted the DNA following the BioSprint 15 DNA Plant Kit protocol (Qiagen Inc.), and eluted the extracted DNA in 200 μ L of AE buffer. Shotgun sequencing (genome skimming approach) was done at the Genopole platform of Toulouse as described in Olofsson et al. [31]. Briefly, 10 to 200 ng of double stranded DNA was used to construct sequencing libraries with the Illumina TruSeq Nano HT Sample kit (Illumina), following the manufacturer's instructions. DNA was fragmented by sonication, except for extracts from herbarium specimens, which were already highly degraded. Each sample was paired-end sequenced (150 bp) on 1/24th of an Illumina HiSeq3000 lane and multiplexed with samples from the same or different projects.

2.2. Assembly of Cytoplasmic and Nuclear DNA Regions

2.2.1. Assembly of Plastome and Nuclear Ribosomal DNA (nrDNA) Cluster

We assembled full plastomes and the nrDNA cluster following the methods of Bianconi et al. [56]. Sequencing depth in these genomic regions was superior to 100 \times for all investigated species. We generated a consensus sequence for both regions for each accession, and mapped reads onto them with GENEIOUS v9.0.5 [57] for manually checking the assembly quality and assessing the sequencing depth. Then, assembled plastomes and nrDNA clusters were annotated in GENEIOUS by transferring annotations from the olive tree (GenBank accessions NC013707.2 and LR031475.1 for plastid and ribosomal cluster, respectively). Finally, we generated independent alignments for the two regions using the MUSCLE algorithm [58] with default options as implemented in GENEIOUS.

2.2.2. Assembly of Mitochondrial Genes

We adopted a reference-based iterative assembly approach to retrieve a set of 37 mitochondrial protein-coding genes for each sampled species (excluding *Olea europaea*, *Capsicum annuum*, and *Solanum lycopersicum*, for which annotated mitochondrial genomes are already available in

GenBank; Table S2). Genes located in regions homologous to plastomes (for which plastid reads mapped on; so called “mtpt” regions) were excluded. Using the reference sequence of the olive tree mitochondrial genes (MG372119.1), an initial set of homologous reads were identified by mapping using Bowtie2 v2.3.5.1 [59] in local mode (all other parameters to default values). These reads were used as the input of a *de novo* assembly using SPAdes v3.14.1 [60] with default parameters. The resulting contigs were then used as reference for the next round of homologous read search and assembly. After three iterations, obtained contigs for each gene were aligned using MAFFT v7.313 [61] with default options. Sequencing depth of mitochondrial genes was superior to 30× for all investigated species. The alignment was then inspected and annotated in GENEIOUS by transferring annotations from the olive tree and extremities were trimmed to the annotated coding-sequence.

2.2.3. Assembly of Genes Encoding Phytochromes

Finally, we analyzed phylogenetic relationships within the Oleaceae using a few nuclear phytochrome genes. Their coding part (cds) is relatively long (>3000 bp; 4 exons) and can be aligned on most of their sequence. A reference-guided approach was used to assemble genomic regions containing genes encoding phytochromes B and E (*phyB*, *phyE*), as described in [62,63]. Briefly, raw genomic data sets were filtered using the NGSQC Toolkit v.2.3.3 [64] to retain only high-quality reads (i.e., >80% of the bases with Phred quality score >20), and to remove adaptor contamination and reads with ambiguous bases. The retained reads were subsequently trimmed from the 3' end to remove bases with Phred score <20. We mapped cleaned paired-end reads on references for genes encoding phytochromes B and E using GENEIOUS. First, exons of *phyB* (two genes, see Phylogenetic analyses below) and *phyE* genes of the ash tree (*Fraxinus excelsior*; GenBank accessions LR983955 to LR983957 [40]) were used as seeds to reconstruct full *phyB* and *phyE* genes of 15 Oleaceae accessions for which nuclear genome sequencing depth was superior or equal to 5× (i.e., *Dimetra craibiana*, *Nyctanthes arbor-tristis*, *Abeliophyllum distichum*, *Forsythia × mandschurica*, *Fontanesia fortunei*, *Jasminum didymum*, *Jasminum pauciflorum*, *ChrysoJasminum fruticans*, *Olea europaea* subsp. *laperrinei*, *Noronhia emarginata*, *Ligustrum ovalifolium*, *Syringa pubescens*, *Schrebera swietenoides*, *Comoranthus obconicus*, and *Fraxinus ornus*). These species are representative of all main Oleaceae lineages (tribes and subtribes) as defined by Wallander and Albert [3]. We carefully checked that *phy* sequences were not chimeric between related paralogs (especially between *phyB-1a* and *phyB-1b*) by a manual verification of reads phasing on gene assemblies. Then, our newly assembled genes were used to assemble exons in other species by using gene sequences of reference from the same tribe or subtribe. Partial or complete consensus coding sequences of *phyB* and *phyE* were thus obtained for the remaining 46 Oleaceae species. Consensus *phy* sequences of *Ny. arbor-tristis* and *Ch. ligustrinus* showed a relatively high rate of ambiguities on all genes [on average 2.38% (2.26–2.50%) and 2.11% (1.6–2.7%), respectively]. A manual checking of these gene assemblies reveals the presence of more than two distinct homologs suggesting we collapsed sequences of recently duplicated genes on these species. Finally, a few paralogs with lower homology to our references were also detected in some accessions and were further considered when their assembly covered more than 1000 bp of the coding sequence. These additional (pseudo)genes were assembled in nine distantly related species (i.e., *Nor. emarginata*, *Chionanthus rupicolus*, *Ch. trichotomus*, *Fore. angustifolia*, *Sc. swietenoides*, *J. didymum*, *A. distichum*, *Fors. mandschurica*, and *Fon. fortunei*). Gene sequences covering more than 90% of the coding region were annotated and deposited in GenBank (Table S3). Genes were considered as potentially non-functional when coding sequences were truncated or presented in-frame stop codon.

2.3. Phylogenetic Analyses

2.3.1. Phylogeny of Oleaceae Using Organellar DNA

All protein-coding sequences were extracted from the full plastomes and aligned separately as codons using PRANK v170427 [65] (default options for translated alignments of protein-coding DNA sequences). We then estimated a tree using the maximum likelihood (ML) algorithm in IQ-Tree2

v2.0.6 [66]. We used a concatenation approach with an edge-linked proportional partition model, using ModelFinder [67], and assessed branch support with 1000 ultrafast bootstrap (UFB) replicates [68]. The best partition scheme for each dataset was determined with PartitionFinder v2.1.1 [69] and the best fitted evolutionary model for each partition was selected according to the best BIC score with ModelFinder, as implemented in IQ-Tree2. An ML phylogenetic tree for the mitochondrial alignment was also estimated, as described above.

2.3.2. Phylogeny of Oleaceae Using nrDNA

In previous studies on the Oleaceae tribe, the nrDNA cluster rendered questionable results with the unexpected phylogenetic clustering of tropical lineages (e.g., Schreberinae subtribe embedded in an Oleinae lineage including genera *Chionanthus*, *Priogymanthus*, *Haenianthus*, *Noronhia*, and *Olea* [30,31,46]). A strong phylogenetic bias was attributed to the highly variable GC content in the external and internal transcribed spacers (ETS and ITS) of the Oleaceae tribe [31] and nrDNA was thus deemed unreliable for phylogenetic inference in this group. However, it has been suggested that a purine-pyrimidine only coding (usually referred to as RY-coding) can effectively reduce the influence of biased GC-content [70]. Before using the nrDNA dataset on the phylogenetic analyses, we thus transformed the data from regular nucleotide-coding to a RY-coding alignment. An ML phylogenetic tree was finally estimated as described above splitting the ribosomal cluster into seven partitions: 5'ETS, 18S, ITS1, 5.8S, ITS2, 26S, and 3'ETS.

2.3.3. Phylogenetic Analyses of the Nuclear *phy* Gene Family

Coding regions of all *phy* sequences were aligned together in a matrix using MAFFT (alignment provided in Supplementary Materials). We then estimated a tree for the *phyB+phyE* gene family by using the ML algorithm in IQ-Tree2. In this case, we estimated the best substitution model for the whole region using ModelFinder [67], and assessed branch support with 1000 UFB replicates. This analysis allowed us to infer ancestral duplications involved in the diversification of the gene family, and then identify orthologs that can be used for reconstructing phylogeny of Oleaceae. Two nuclear genes (*phyB-1* and *phyE-1*), putatively encoding functional enzymes in most analyzed accessions, were finally selected for the phylogenetic inference of the Oleaceae family. In Oleaceae, two paralogs (*phyB-1a* and *phyB-1b*) were kept, with *phyB-1a* arbitrarily aligned to the *phyB-1* copies of other Oleaceae tribes. An ML phylogenetic tree was finally estimated as described above allowing one partition per gene.

2.3.4. Phylogenetic Inference of Family Tree Using Data from Mixed Origin

We then estimated an ML phylogeny for Oleaceae combining nuclear and organellar information and assessed congruence between the datasets by using the algorithm for concordance factors calculations implemented in IQ-Tree2. We quantified the concordance between this phylogeny and each dataset by calculating the gene concordance factor (gCF) and the site concordance factor (sCF) for each branch of the reference tree [71]. The gCF represents the fraction of individual trees (here, species tree obtained with one of the datasets) that is concordant with a given branch, and the sCF shows the proportion of alignment sites that support that branch. It thus allows us to quantify the presence of sites inside each dataset supporting the combined topology, even if the topology obtained with one individual dataset shows an alternative topology.

3. Results

3.1. Phylogenetic Reconstructions Based on Chloroplast and Mitochondrial Genes

Using chloroplastic gene data (consisting of 77,676 sites including 10,059 parsimony-informative sites), we obtained a fully-resolved tree of the family (Figure 1). Oleaceae division into five tribes (Myxopyreae, Jasmineae, Forsythieae, Fontanesieae, and Oleaceae) is strongly supported. In this dataset, Myxopyreae forms a monophyletic tribe (with the *Myxopyrum* genus sister to *Dimetra+Nyctanthes*)

and is the sister lineage to all other groups in Oleaceae. Jasmineae appears as sister group to Oleaeae. Schreberinae are represented as the sister clade (and subtribe) to the rest of the clades in the monophyletic tribe Oleaeae, and *Schrebera* is paraphyletic. Within the Oleaeae subtribe Ligustrinae, the genus *Syringa* also forms a paraphyletic group. Within Oleinae, the tree consists of short branches with a few polyphyletic genera (i.e., *Chionanthus*, *Olea*, *Osmanthus*, and *Nestegis*). Branch length was particularly long in tribe Jasmineae (notably in *Menodora*) and at a lesser extent in the core Ligustrinae and *Dimetra*+*Nyctanthes*, suggesting an increase of the evolutionary rate of plastid genes in these clades.

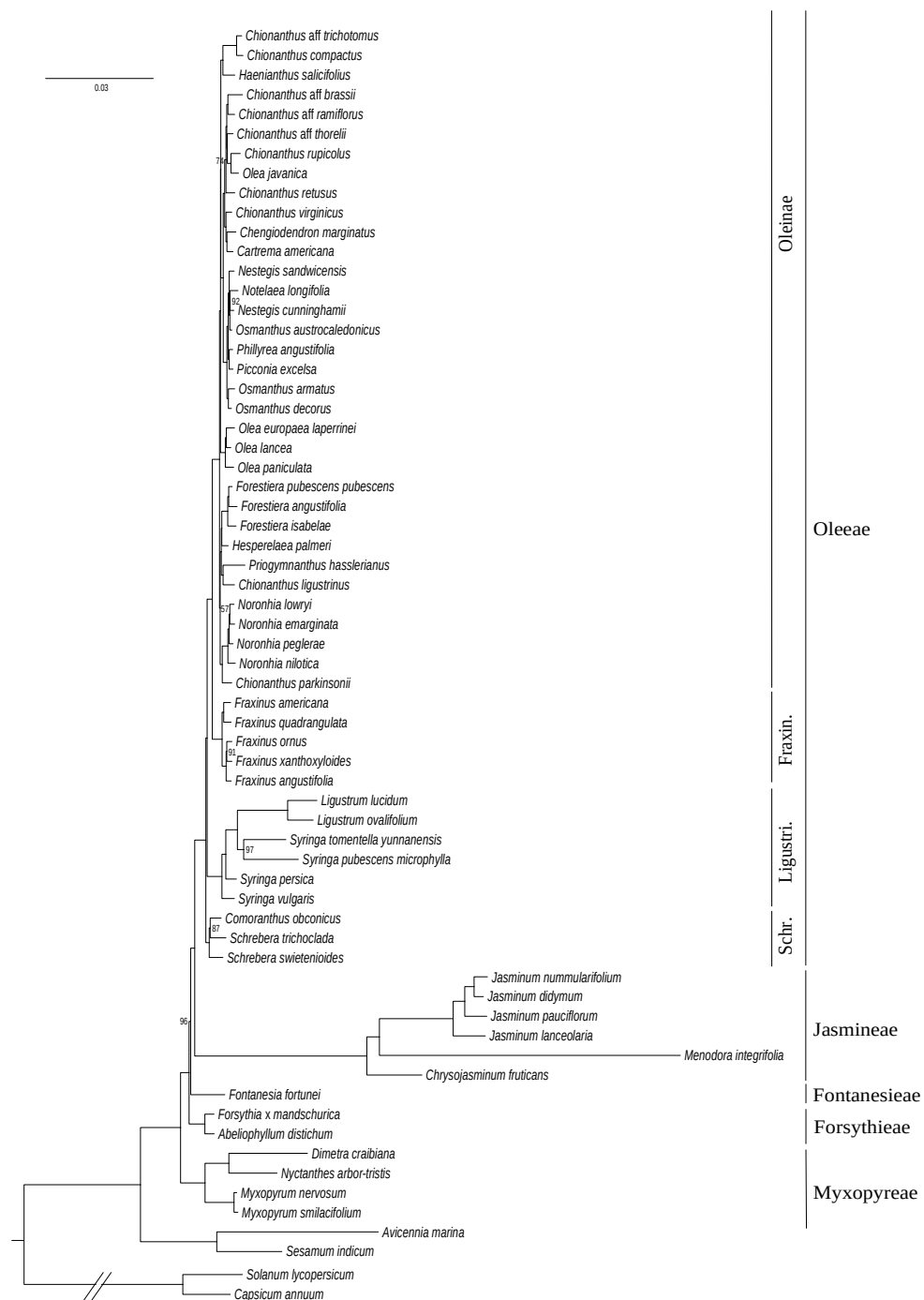


Figure 1. Maximum likelihood phylogenetic tree of Oleaceae based on concatenated coding sequences of 80 plastid genes. The tree was rooted on the split with Solanaceae. The scale is in substitution per site. Ultrafast bootstrap support values are indicated on nodes when inferior to 100.

In comparison to the chloroplastic DNA phylogeny, the phylogeny based on mitochondrial data (60,747 sites, 3509 parsimony-informative sites) exhibits a highly-congruent albeit less supported topology (Figure 2). We only stress one significant difference, regarding the branching order in the deepest nodes of the family, in this topology, Forsythieae is positioned as the sister clade to all other Oleaceae (and not Myxopyreae as in the chloroplast tree). Again, Jasmineae (especially *Menodora*) and *Dimetra*+*Nyctanthes* show longer branches suggesting an increase of the evolutionary rate in these two clades.

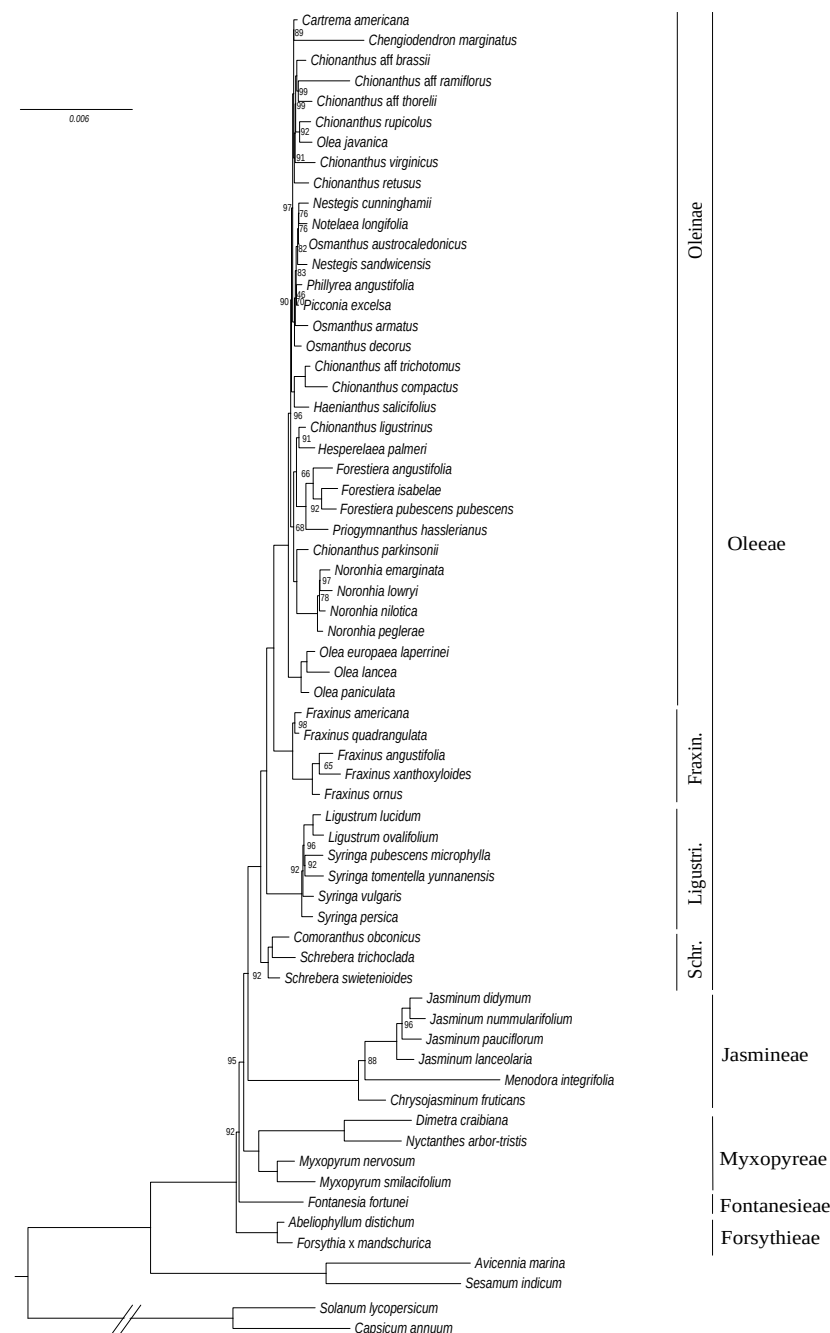


Figure 2. Maximum likelihood phylogenetic tree of Oleaceae based on the concatenation of 37 mitochondrial genes. The tree was rooted on the split with Solanaceae. The scale is in substitution per site. Ultrafast bootstrap support values are indicated on nodes when inferior to 100.

3.2. Phylogeny Based on the Nuclear Ribosomal Cluster

Compared to phylogenetic reconstructions based on cytoplasmic genes, the analysis of the nrDNA cluster (7008 sites, 837 parsimony-informative sites) resulted in a less-supported and quite different topology (Figure 3). Myxopyreae+Fontanesieae+Forsythieae are resolved as sister to the tribes Jasmineae and Oleae. Myxopyreae are not monophyletic, with *Myxopyrum* sister to *Forsythia*+*Fontanesia* but this topology is poorly supported (UFB:64). Jasmineae is here again reported as sister to Oleae but includes a different branching of *Menodora* (sister to *Jasminum*+*Chrysojasminum*). This topology presents a first strongly-supported split in Oleae between Schreberinae and Fraxininae+Ligustrinae+Oleinae (UFB:97). Within this grouping, Fraxininae and Ligustrinae form monophyletic lineages sister to Oleinae but are not supported. Longer branches are still observed in Jasmineae and *Dimetra*+*Nyctanthes* (especially in *Dimetra*).

3.3. Phylogeny Based on Nuclear *phy* Gene Family

A second nuclear DNA phylogeny was reconstructed using *phy* genes. We first investigated the phylogenetic tree of the *phy* family in order to select the most informative orthologs. A condensed phylogenetic tree of genes encoding phytochromes E and B is shown in Figure 4 (the detailed tree is provided in Figure S1). As expected, the main distinction of two genes, *phyE* and *phyB*, was recovered.

For *phyE*, one supposedly functional gene (*phyE-1*) was detected in most Oleaceae species, although a second functional gene (*phyE-2*) was also assembled in tribes Forsythieae and Fontanesieae. *phyE-2* is sister to a clade formed by *phyE-1* and *phyE* of *Avicennia* and *Sesamum* (recovered from GenBank). This topology suggests an ancestral gene duplication (giving birth to *phyE-1* and *phyE-2*) in the ancestor of Lamiales, after its divergence from Solanales. A likely pseudogenetic *phyE-1* paralog (namely *phyE-1b*) was detected in *Schrebera swietenoides* (Oleae). Its phylogenetic position remains unresolved due to a polytomy with *phyE-1* clades of Oleae (namely *phyE-1a*) and Jasmineae. *phyE-1b* likely testifies to a gene duplication in the Oleae ancestor [3,4], followed by a rapid pseudogenization of this duplicate. Interestingly, we also detected putative pseudogenes of *phyE-2* in distantly related species of Jasmineae and Oleae. Two putatively pseudogenetic lineages were detected in Oleae (*phyE-2a* and *phyE-2b*), another evidence of (pseudo)gene duplication in the ancestor of this tribe [3,4]. Based on this topology, only *phyE-1* was selected for our phylogenetic analyses of species relationships because this ortholog was detected in all analyzed Oleaceae accessions, and phylogenetic relationships based on this gene support the main taxonomic lineages (i.e., tribes and subtribes) as defined by Wallander and Albert [3]. Putatively pseudogenized copies (i.e., presence of frame shifts and/or stop codons) of *phyE-1a* were detected in eight species (Figure S1).

For *phyB*, first, two functional duplicates were detected in Solanales, Acanthaceae (*Avicennia*) and Pedaliaceae (*Sesamum*). Two main gene lineages (*phyB-1* and *phyB-2*) were also detected in Oleaceae, but *phyB-2* was detected only in Forsythieae (*Forsythia* and *Abeliophyllum*). This gene is sister to the *phyB* genes of Acanthaceae and Pedaliaceae. On the other hand, *phyB-1* was detected in all Oleaceae species. Two closely related genes (*phyB-1a* and *phyB-1b*) were assembled in all Oleae species, again testifying to an event of gene duplication in the ancestor of this tribe [3,4]. Based on this topology, *phyB-1* was selected for species relationships analyses because this gene was detected in all analyzed accessions, and the phylogeny allowed us to retrieve all Oleaceae lineages [3]. Putatively pseudogenetic copies (i.e., presence of frame shifts and/or stop codons or complete deletion of exon) of *phyB-1a* and *phyB-1b* were detected in two and four species, respectively (Figure S1).

The phylogenetic tree based on concatenated *phyB-1* (*a* and *b*) and *phyE-1* genes (10,438 sites, 3282 parsimony-informative sites) is shown in Figure 5. Again, the topology supports the distinction of all taxonomic units defined by Wallander and Albert [3], with tribe Myxopyreae recognized as sister to the rest of Oleaceae. As in other topologies showed above, tribes Jasmineae and Oleae as well as subtribes Oleinae and Fraxininae are sister groups. In contrast, a major incongruence with both cytoplasmic datasets is the placement of subtribe Ligustrinae as sister to the remaining of

Oleaceae. This topology was recovered with *phyB-1a* and *phyE-1a*, but not with *phyB-1b* that supports Schreberinae as sister to the other subtribes (Figures 4 and S1). Longer branches are observed in Jasmineae and *Dimetra*.

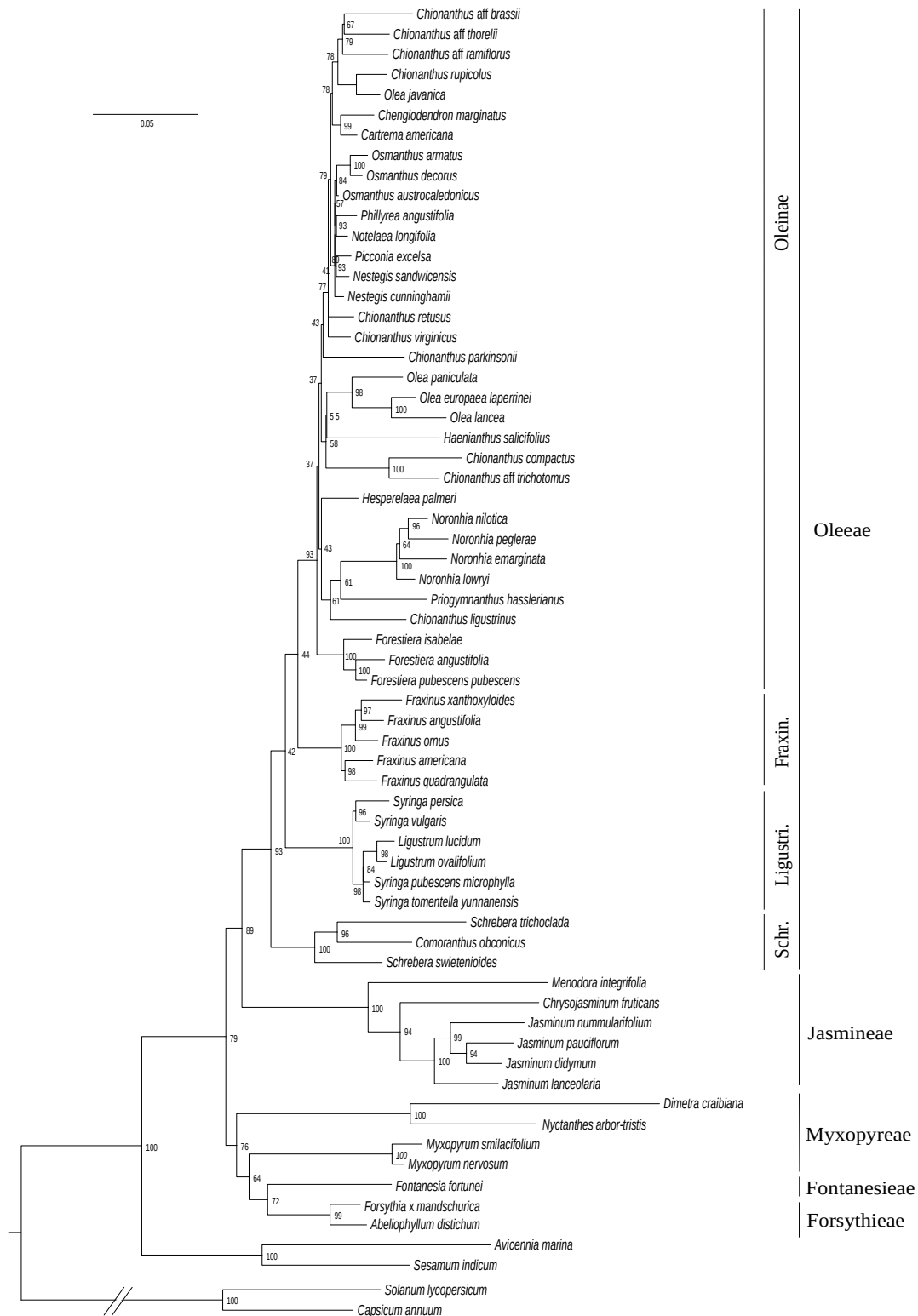


Figure 3. Maximum likelihood phylogenetic tree of Oleaceae based on the complete RY-coded nrDNA cluster alignment. The tree was rooted on the split with Solanaceae. The scale is in substitution per site.

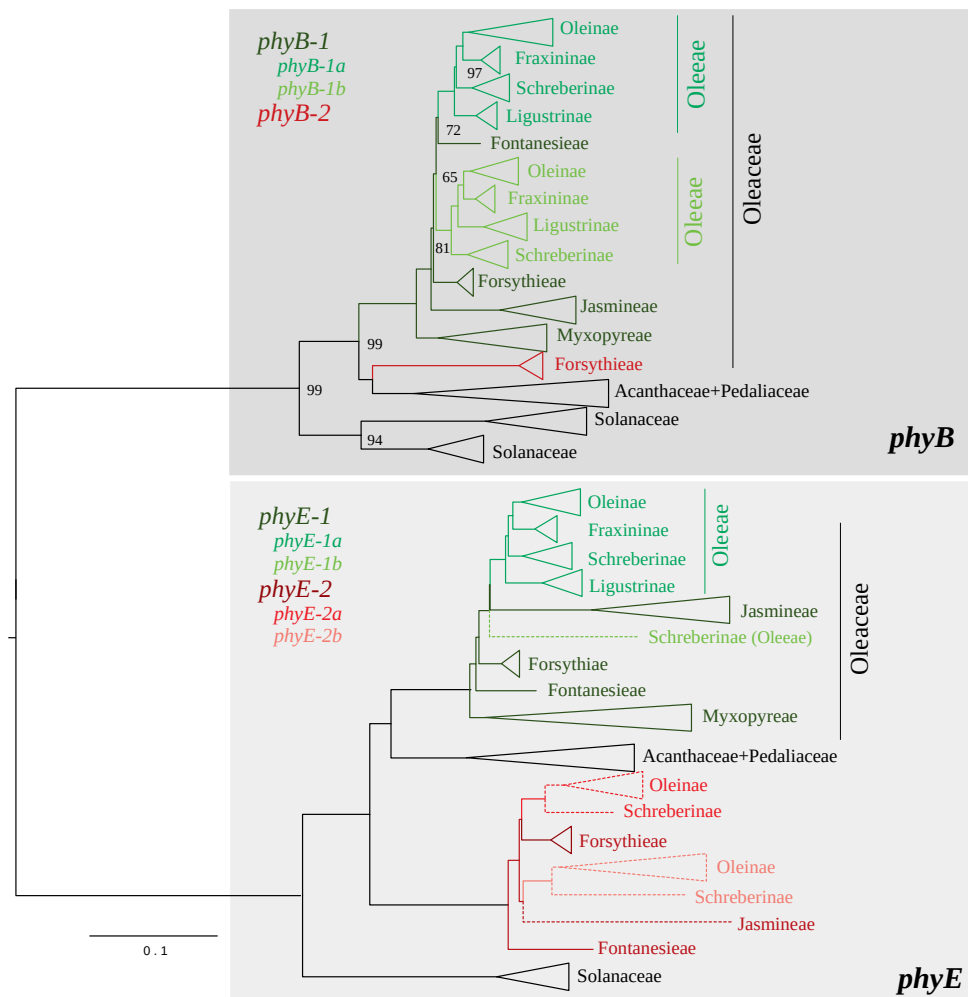


Figure 4. Reduced representation of the midpoint-rooted maximum likelihood phylogenetic tree of the *phy* gene family in Oleaceae. Only ultrafast bootstrap (UFB) values inferior to 100 are indicated on nodes. Putative pseudogenes are denoted by dashed lines.

3.4. Phylogenetic Reconstruction Combining the Four Genomic Datasets

The combination of nuclear and cytoplasmic datasets allowed the reconstruction of a well-supported phylogeny of Oleaceae (Figure 6). All datasets broadly supported the same phylogenetic hypothesis with five strongly supported monophyletic tribes Myxopyreae, Fontanesieae, Forsythieae, Jasmineae and Oleaeae. The position of Myxopyreae as sister to the rest of the family is supported by the majority of data as concordance factors attest. The branching order of Forsythieae and Fontanesieae is however difficult to decide on. For these two tribes, the topology of the species tree obtained from the combined dataset is not well-supported. The branching node of Forsythieae, despite a bootstrap support of 100, exhibits high uncertainty based on the concordance factors (gCF: 50%; sCF: 51.4%, Figure S2). The represented branching of Fontanesieae is even less supported (UFB: 64; gCF: 25%; sCF: 29.1%, Figure S2). In both cases, concordance factors show that the reported topology is not supported by most sites. Similar sCF and gCF values suggest this is due to genuine discordant signal in the trees probably due to incomplete lineage sorting. In contrast, we set Jasmineae as the sister tribe of Oleaeae with confidence (UFB and gCF values of 100). The topology within Jasmineae confirms the recent reevaluation of the genus *Jasminum* in two distinct genera *Chrysojasminum* and *Jasminum* [36,37,54]. The other major uncertainty resides within the Oleaeae tribe on the branching

order of Ligustrinae and Schreberinae. Although bootstrap support and concordance factors values sustain the represented branching (Schreberinae as sister to other Oleaceae subtribes), the concordance factors (especially sCF) are less decisive for the Ligustrinae split.

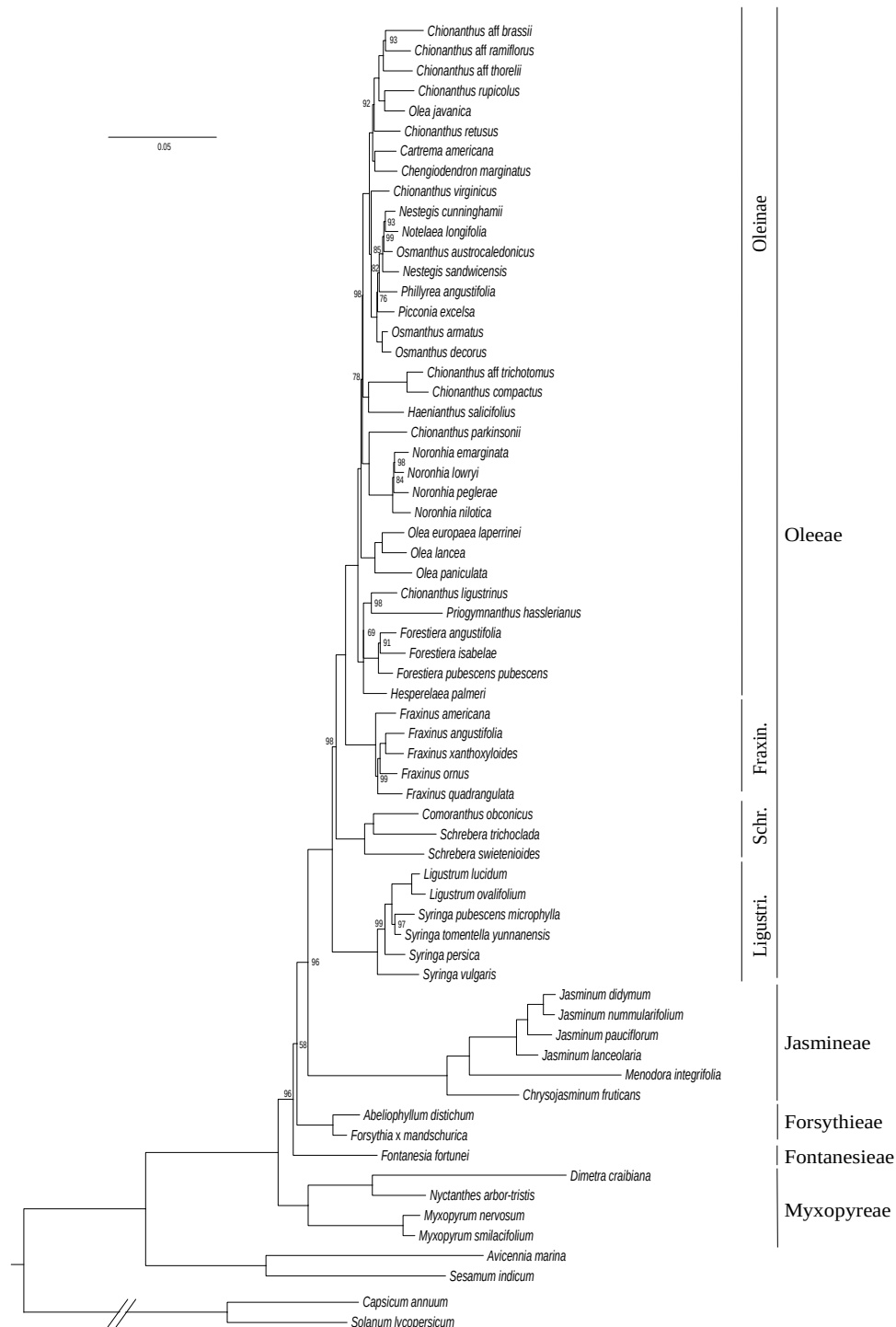


Figure 5. Maximum likelihood phylogenetic tree of Oleaceae based on *phyB-1* (a and b) and *phyE-1* nuclear genes. Oleaceae *phyB-1a* was arbitrarily aligned with *phyB-1* of other tribes. The tree was rooted on the split with Solanaceae. The scale is in substitution per site. Ultrafast bootstrap support values are indicated on nodes when inferior to 100.

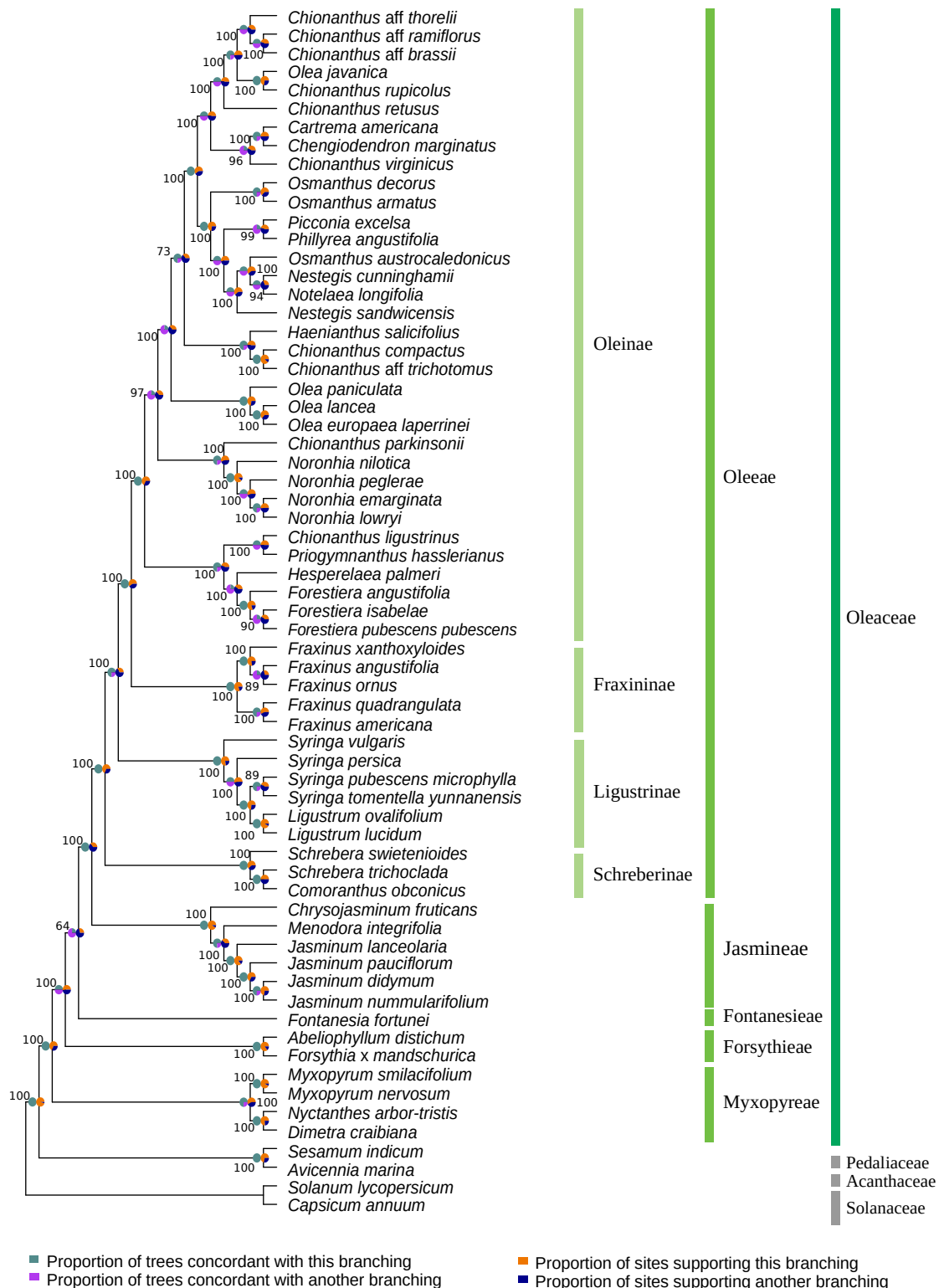


Figure 6. Maximum likelihood topology of Oleaceae family estimated from the partitioned concatenation of 80 plastid coding sequence, 37 mitochondrial genes, the complete nuclear ribosomal DNA cluster and three nuclear genes encoding phytochromes (*phyE-1*, *phyB-1a*, *phyB-1b*). Concordance factors were calculated in relation to the species trees inferred for each partitioned dataset. Gene concordance factors are represented by the green/purple pie charts (left), site concordance factors by the blue/orange ones (right). UFB support values are indicated near their respective nodes.

4. Discussion

We gathered molecular information from several genomic compartments (chloroplastic, mitochondrial and nuclear) for 61 Oleaceae species representative of all currently recognized tribes, subtribes and genera in Oleaceae. Both plastid and mitochondrial DNA datasets as well as the nrDNA cluster are based on relatively high sequencing depth (>30×) and thus of a high quality [30,41]. In contrast, low-copy nuclear genes are more difficult to assemble from genome skimming data and their use in phylogenetics is still a challenge due to lower coverage and recurrent whole genome duplications [31,48]. Here, we explored the utility of a single nuclear gene family (*phyB* and *phyE* genes) for investigating the phylogeny of the whole Oleaceae family. The obtained dataset allowed us to tackle the complex history of nuclear gene duplication and subsequent pseudogenization indicating the necessity to control for gene orthology before proposing a phylogenetic hypothesis for the whole family. By combining and confronting our datasets, we were able to establish a well-resolved phylogeny of Oleaceae although a few discordances were revealed when comparing phylogenies based on cytoplasmic and nuclear genomic regions. Overall, tribes and subtribes were strongly supported by all phylogenetic reconstructions and only very few relationships between tribes/subtribes were not fully resolved.

4.1. Taxonomy of Oleaceae

Our phylogenetic analyses confirm the divisions of Oleaceae in five tribes and four subtribes as defined by Wallander and Albert [3]. Given the amount of data we analyzed, we achieved a greater resolution and support in our phylogenetic inference of the whole family, including all currently recognized genera and considering several accessions from distant areas in the largest groups (e.g., *Chionanthus*, *Olea*, *Fraxinus*, *Syringa*, *Jasminum*). First, our results validated the grouping of *Nyctanthes*, *Dimetra* and *Myxopyrum* in Myxopyreae [3,72] and overall supported this clade as sister to all other lineages in the family. We were also able to corroborate some of the less-reliable nodes and in particular the sister tribes Jasmineae and Oleae. We resolved the relationships between Forsythiae and Fontanesieae as being distinct and non-sister tribes. We also put into question the idea that Ligustrinae is sister to all other lineages in Oleae [3,35,36] favoring the alternative hypothesis of Schreberinae being the one (as in [31], where the whole plastid genome and single-nucleotide polymorphisms datasets gathered from more than 11,000 nuclear genes were used). Finally, we were also able to better define the relationships within Oleae wherein some genera appeared as polyphyletic (i.e., *Chionanthus*, *Olea*, *Osmanthus*, *Nestegis*) or paraphyletic (i.e., *Schrebera*, *Syringa*) confirming previous reports from the literature [26,28,30–32].

A relatively high congruence was obtained between phylogenies based on plastid and mitochondrial DNA datasets (Figures 1 and 2), as expected for maternally inherited genomes [42]. We obtained the best resolution with the chloroplastic dataset as it contains more informative sites. Topologies based on *phy* genes and cytoplasmic genomes were also quite congruent although the relative placement of Ligustrinae and Schreberinae as well as Forsythiae and Fontanesieae differ according to *phy* genes (Figures 4 and 5). In contrast, the nrDNA cluster provided less reliable information than organellar genomes and *phy* nuclear genes (Figure 3). Phylogenetic biases related to GC content and incomplete concerted evolution have been already reported in Oleaceae for the nrDNA marker (e.g., [10,31,46]), which thus needs to be interpreted with caution. Yet, the RY-coding seems to have greatly improved the topology since all Oleae subtribes were retrieved in contrast to previous analyses [31,46] (see Figure S3 for the ML phylogeny from the original alignment).

4.2. Nuclear Gene Orthology and Polyploidization Events in Oleaceae

The analysis of a small multigene family revealed other aspects on the Oleaceae history, related to past whole genome duplications and different tempo of pseudogenization. First, two divergent functional paralogs were revealed on *phyE* and *phyB*, but only in Fontanesieae and Forsythiae.

The duplication of these genes (possibly due to whole genome duplication) is ancient, likely preceding the divergence of Lamiales, and the pseudogenisation of *phy-B2* and *phy-E2* in tribes Myxopyreae, Jasmineae and Oleaeae may have occurred rapidly after their divergence. Only pseudo-*phy-E2* was still detected in Jasmineae and Oleaeae. More interestingly, the detection of two closely related paralogs of *phyB-1*, *phyE-1* and pseudo-*phyE-2* in all Oleaeae species is highly congruent with the reported event of polyploidization in their common ancestor [3,4]. As we decided to collapse highly homologous sequences of *phy* genes, we were not able to investigate the fate of these genes in neopolyploids, but we detected a relatively high level of ambiguities in the tetraploid *Ny. arbor-tristis* [73] as well as in *Ch. ligustrinus* for which the chromosome number is unknown.

5. Concluding Remarks and Future Directions in Oleaceae Phylogenomics

Our work provided a more robust phylogenetic history of Oleaceae than previous works, a crucial prerequisite to study the diversification process of this family. A complex history of gene duplication and pseudogenization was also revealed, and these aspects need to be evaluated before using nuclear data in the reconstruction of phylogenies, especially in a plant family with paleopolyploids such as Oleaceae. Moreover, our prospective study also demonstrated the limits of using *phy* genes to estimate a tree due to the variable levels of gene retention and the presence of non-functional sequences. With the higher accessibility of genomic data, some of these caveats can be circumvented with the use of new methodologies such as the analyses of UCE (Ultra Conserved Elements) or universal single-copy orthologs (e.g., [74–76]). Although, in the light of the complicated history of evolution of plants (e.g., multiple reported events of whole genome duplication), we stress the importance of taking gene orthology into account when estimating species trees.

When it comes to our current and future goals with the study of the phylogenomics of Oleaceae, the complete sequencing of nuclear genomes (with at least 30–50× coverage) is in progress in our lab. We are mainly focusing on low heterozygous diploid species, and avoiding neo-polyploids and hybrids. In addition, since this study confirmed that cytoplasmic and nuclear ribosomal DNA sequences can be easily assembled independent of species ploidy, we are using those genomic regions on a comprehensive sampling to reconstruct a fossil-calibrated phylogeny of the family. Finally, with this large phylogeny of Oleaceae we will explore the causes of variable evolutionary rates among genomes, considering factors as generation time (e.g., short living species exhibit particularly long branches in phylogenetic reconstructions) [77], gene duplication, genome inheritance, and recombination rate [77–79].

Supplementary Materials: The following are available online at <http://www.mdpi.com/2073-4425/11/12/1508/s1>. Table S1. List of Oleaceae accessions analyzed in our study, with their taxonomy, accession number and origin. Table S2. List of species used as outgroups in our phylogenetic analyses. Table S3. GenBank no of genomic regions for each accession. Figure S1. Full representation of the midpoint-rooted maximum likelihood phylogenetic tree of the *phy* gene family in Oleaceae. Figure S2. Maximum likelihood topology of Oleaceae estimated from the partitioned analysis of the four datasets with corresponding concordance factors of nodes. Figure S3. Maximum likelihood phylogenetic tree of Oleaceae based on the non-transformed nrDNA cluster alignment. Materials S1 to S11. Sequence alignments used for phylogenetic reconstructions, and tree files.

Author Contributions: Conceptualization: J.D. and G.B.; Plant sampling: J.D., C.H.-W., M.G., and G.B.; Lab work: J.D., S.M., and G.B.; Data analyses: J.D., P.R., and G.B.; Manuscript writing: J.D., P.R., and G.B.; Funding acquisition: J.D. and G.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Fruitful grant (H2020-MSCA-IF-2018-842234), the ERA-NET BiodivERsA project INFRAGECO (ANR-16-EBI3-0014), and by the grant GeneRes (Occitanie-France Olive). In addition, J.D., P.R., S.M. and G.B. are members of the EDB laboratory, which is supported by the excellence projects Labex CEBA (ANR-10-LABX-25-01) and Labex TULIP (ANR-10-LABX-0041), managed by the French ANR.

Acknowledgments: We are grateful to the Genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, doi:10.15454/1.5572369328961167E12) for providing computing and storage resources, and to Céline Van de Paer for help in gene annotation of plastomes. We also thank three anonymous reviewers for their constructive comments.

Conflicts of Interest: The authors declare no competing interests. Funders played no role in the study design; data collection, analysis or interpretation; in the writing of the manuscript; or in the decision to publish the results.

References

1. Green, P.S. Oleaceae. In *The Families and Genera of Vascular Plants, Flowering Plants, Dicotyledons*; Kubitzki, K., Kadereit, J.W., Eds.; Springer: New York, NY, USA, 2004; Volume 7, pp. 296–306.
2. Mitchell, R.J.; Beaton, J.K.; Bellamy, P.E.; Broome, A.; Chetcuti, J.; Eaton, S.; Ellis, C.J.; Gimona, A.; Harmer, R.; Hester, A.J.; et al. Ash dieback in the UK: A review of the ecological and conservation implications and potential management options. *Biol. Conserv.* **2014**, *175*, 95–109. [[CrossRef](#)]
3. Wallander, E.; Albert, V.A. Phylogeny and classification of Oleaceae based on *rps16* and *trnL-F* sequence data. *Am. J. Bot.* **2000**, *87*, 1827–1841. [[CrossRef](#)] [[PubMed](#)]
4. Taylor, H. Cyto-taxonomy and phylogeny of the Oleaceae. *Brittonia* **1945**, *5*, 337–367. [[CrossRef](#)]
5. Briggs, B.G. Some chromosome numbers in the Oleaceae. *Contrib. N. S. W. Natl. Herb.* **1970**, *4*, 126–129.
6. George, K.; Geethamma, S. Cytology and evolution of jasmines. *Cytologia* **1992**, *57*, 27–32. [[CrossRef](#)]
7. Besnard, G.; Garcia-Verdugo, C.; Rubio de Casas, R.; Treier, U.A.; Galland, N.; Vargas, P. Polyploidy in the olive complex (*Olea europaea*): Evidence from flow cytometry and nuclear microsatellite analyses. *Ann. Bot.* **2008**, *101*, 25–30. [[CrossRef](#)]
8. Lattier, J.D.; Contreras, R.N. Ploidy and genome size in lilac species, cultivars, and interploid hybrids. *J. Am. Soc. Hortic. Sci.* **2017**, *142*, 355–366. [[CrossRef](#)]
9. Whittemore, A.T.; Cambell, J.J.N.; Zheng-Lian, X.; Carlson, C.H.; Atha, D.; Olsen, R.T. Ploidy variation in *Fraxinus* L. (Oleaceae) of eastern North America: Genome size diversity and taxonomy in a suddenly endangered genus. *Int. J. Plant Sci.* **2018**, *179*, 377–389. [[CrossRef](#)]
10. Wallander, E. Systematics of *Fraxinus* (Oleaceae) and evolution of dioecy. *Plant Syst. Evol.* **2008**, *273*, 25–49. [[CrossRef](#)]
11. Hinsinger, D.D.; Bask, J.; Gaudeul, M.; Cruaud, C.; Bertolino, P.; Frascaria-Lacoste, N.; Bousquet, J. The phylogeny and biogeographic history of ashes (*Fraxinus*, Oleaceae) highlight the roles of migration and vicariance in the diversification of temperate trees. *PLoS ONE* **2013**, *8*, e80431. [[CrossRef](#)]
12. Rohwer, J.G. A preliminary survey of the fruits and seeds of the Oleaceae. *Bot. Jahrb. Syst.* **1993**, *115*, 271–291.
13. Rohwer, J.G. Fruit and seed structures in *Menodora* (Oleaceae): a comparison with *Jasminum*. *Bot. Acta* **1995**, *108*, 163–168. [[CrossRef](#)]
14. Rohwer, J.G. Die Frucht- und Samenstrukturen der Oleaceae. *Bibl. Bot.* **1996**, *148*, 1–177.
15. Kiew, R. Preliminary pollen study of the Oleaceae in Malesia. *Gard. Bull.* **1984**, *37*, 225–230.
16. Lepart, J.; Dommée, B. Is *Phillyrea angustifolia* L. (Oleaceae) an androdioecious species? *Bot. J. Linn. Soc.* **1992**, *108*, 375–387. [[CrossRef](#)]
17. Green, P.S. A revision of *Olea* L. (Oleaceae). *Kew Bull.* **2002**, *57*, 91–140. [[CrossRef](#)]
18. Saumitou-Laprade, P.; Vernet, P.; Dowkiw, A.; Bertrand, S.; Billiard, S.; Albert, B.; Gouyon, P.H.; Dufay, M. Polygamy or subdioecy? The impact of diallelic self-incompatibility on the sexual system in *Fraxinus excelsior* (Oleaceae). *Proc. R. Soc. B Biol. Sci.* **2018**, *285*, 20180004. [[CrossRef](#)]
19. Thompson, J.D.; Dommée, B. Morph-specific patterns of variation in stigma height in natural populations of distylous *Jasminum fruticans*. *New Phytol.* **2000**, *148*, 303–314. [[CrossRef](#)]
20. Olesen, J.M.; Dupont, Y.L.; Ehlers, B.K.; Valido, A.; Hansen, D.M. Heterostyly in the Canarian endemic *Jasminum odoratissimum* (Oleaceae). *Nord. J. Bot.* **2005**, *23*, 537–539. [[CrossRef](#)]
21. Ryu, T.Y.; Yeom, D.Y.; Kim, Y.J.; Kim, S.J. Studies on heterostyly incompatibility of *Abeliophyllum distichum*. *Seoul Natl. Univ. Coll. Agric. Bull.* **1976**, *1*, 113–120.
22. Saumitou-Laprade, P.; Vernet, P.; Vassiliadis, C.; Hoareau, Y.; de Magny, G.; Dommée, B.; Lepart, J. A self-incompatibility system explains high male frequencies in an androdioecious plant. *Science* **2010**, *327*, 1648–1650. [[CrossRef](#)] [[PubMed](#)]
23. Vernet, P.; Lepercq, P.; Billiard, S.; Bourceaux, A.; Lepart, J.; Dommée, B.; Saumitou-Laprade, P. Evidence for the long-term maintenance of a rare self-incompatibility system in Oleaceae. *New Phytol.* **2016**, *210*, 1408–1417. [[CrossRef](#)] [[PubMed](#)]
24. Johnson, L.A.S. A review of the family Oleaceae. *Contrib. N. S. W. Natl. Herb.* **1957**, *2*, 395–418.
25. Stearn, W.T. Union of *Chionanthus* and *Linociera* (Oleaceae). *Ann. Mo. Bot. Gard.* **1976**, *63*, 355–357. [[CrossRef](#)]

26. Li, J.; Alexander, J.H.; Zhang, D. Paraphyletic *Syringa* (Oleaceae): Evidence from sequences of nuclear ribosomal DNA ITS and ETS regions. *Syst. Bot.* **2002**, *27*, 592–597.
27. Harborne, J.B.; Green, P.S. A chemotaxonomic survey of flavonoids in leaves of the Oleaceae. *Bot. J. Linn. Soc.* **1980**, *81*, 155–167. [[CrossRef](#)]
28. Besnard, G.; Rubio de Casas, R.; Christin, P.A.; Vargas, P. Phylogenetics of *Olea* (Oleaceae) based on plastid and nuclear ribosomal DNA sequences: Tertiary climatic shifts and lineage differentiation times. *Ann. Bot.* **2009**, *104*, 143–160. [[CrossRef](#)]
29. Yuan, W.J.; Zhang, W.R.; Han, Y.J.; Dong, M.F.; Shang, F.D. Molecular phylogeny of *Osmanthus* (Oleaceae) based on non-coding chloroplast and nuclear ribosomal internal transcribed spacer regions. *J. Syst. Evol.* **2010**, *48*, 482–489. [[CrossRef](#)]
30. Hong-Wa, C.; Besnard, G. Intricate patterns of phylogenetic relationships in the olive family as inferred from multi-locus plastid and nuclear DNA sequence analyses: A close-up on *Chionanthus* and *Noronhia* (Oleaceae). *Mol. Phylogenet. Evol.* **2013**, *67*, 367–378. [[CrossRef](#)]
31. Olofsson, J.K.; Cantera, I.; Van de Paer, C.; Hong-Wa, C.; Zedane, L.; Dunning, L.T.; Alberti, A.; Christin, P.A.; Besnard, G. Phylogenomics using low-depth whole genome sequencing: A case study with the olive tribe. *Mol. Ecol. Resour.* **2019**, *19*, 877–892. [[CrossRef](#)]
32. Li, Y.F.; Zhang, M.; Wang, X.R.; Sylvester, S.P.; Xiang, Q.B.; Li, X.; Li, M.; Zhu, H.; Zhang, C.; Chen, L.; et al. Revisiting the phylogeny and taxonomy of *Osmanthus* (Oleaceae) including description of the new genus *Chengiodendron*. *Phytotaxa* **2020**, *436*, 283–292. [[CrossRef](#)]
33. Kim, K.J.; Jansen, R.K. A chloroplast DNA phylogeny of lilacs (*Syringa*, Oleaceae): Plastome groups show a strong correlation with crossing groups. *Am. J. Bot.* **1998**, *85*, 1338–1351. [[CrossRef](#)] [[PubMed](#)]
34. Lee, H.L.; Jansen, R.K.; Chumley, T.W.; Kim, K.J. Gene relocations within chloroplast genomes of *Jasminum* and *Menodora* (Oleaceae) are due to multiple, overlapping inversions. *Mol. Biol. Evol.* **2007**, *24*, 1161–1180. [[CrossRef](#)]
35. Kim, D.; Kim, J. Molecular phylogeny of tribe Forsythieae (Oleaceae) based on nuclear ribosomal DNA internal transcribed spacers and plastid DNA *trnL-F* and *matK* gene sequences. *J. Plant Res.* **2011**, *124*, 339–347. [[CrossRef](#)]
36. Ha, Y.H.; Kim, C.; Choi, K.; Kim, J.H. Molecular phylogeny and dating of Forsythieae (Oleaceae) provide insight into the Miocene history of Eurasian temperate shrubs. *Front. Plant Sci.* **2018**, *9*, 99. [[CrossRef](#)] [[PubMed](#)]
37. Jeyarani, J.N.; Yohannan, R.; Vijayavalli, D.; Dwivedi, M.D.; Pandey, A.K. Phylogenetic analysis and evolution of morphological characters in the genus *Jasminum* L. (Oleaceae) in India. *J. Genet.* **2018**, *97*, 1225–1239. [[CrossRef](#)]
38. Cruz, F.; Julca, I.; Gómez-Garrido, J.; Loska, D.; Marcet-Houben, M.; Cano, E.; Galán, B.; Frias, L.; Ribeca, P.; Derdak, S.; et al. Genome sequence of the olive tree, *Olea europaea*. *GigaScience* **2016**, *5*, 29. [[CrossRef](#)]
39. Unver, T.; Wu, Z.; Sterck, L.; Turktas, M.; Lohaus, R.; Li, Z.; Yang, M.; He, L.; Deng, T.; Escalante, F.J.; et al. Genome of wild olive and the evolution of oil biosynthesis. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E9413–E9422. [[CrossRef](#)]
40. Sollars, E.S.A.; Harper, A.L.; Kelly, L.J.; Sambles, C.M.; Ramirez-Gonzalez, R.H.; Swarbreck, D.; Kaithakottil, G.; Cooper, E.D.; Uauy, C.; Havlickova, L.; et al. Genome sequence and genetic diversity of European ash trees. *Nature* **2017**, *541*, 212–216. [[CrossRef](#)]
41. Kelly, L.J.; Plumb, W.J.; Carey, D.W.; Mason, M.E.; Cooper, E.D.; Crowther, W.; Whittemore, A.T.; Rossiter, S.J.; Kock, J.L.; Buggs, R.J.A. Convergent molecular evolution among ash species resistant to the emerald ash borer. *Nat. Ecol. Evol.* **2020**, *4*, 1116–1128. [[CrossRef](#)]
42. Van de Paer, C.; Bouchez, O.; Besnard, G. Prospects on the evolutionary mitogenomics of plants: A case study on the olive family (Oleaceae). *Mol. Ecol. Resour.* **2018**, *18*, 409–423. [[CrossRef](#)] [[PubMed](#)]
43. Zhang, C.; Zhang, T.; Luebert, F.; Xiang, Y.; Huang, C.H.; Hu, Y.; Rees, M.; Frohlich, M.W.; Qi, J.; Weigend, M.; et al. Asterid phylogenomics/phylotranscriptomics uncover morphological evolutionary histories and support phylogenetic placement for numerous whole-genome duplications. *Mol. Biol. Evol.* **2020**, *37*, 3188–3210. [[CrossRef](#)] [[PubMed](#)]
44. Bieker, V.C.; Martin, M.D. Implications and future prospects for evolutionary analyses of DNA in historical herbarium collections. *Bot. Lett.* **2018**, *165*, 409–418. [[CrossRef](#)]

45. Van de Paer, C.; Hong-Wa, C.; Jeziorski, C.; Besnard, G. Mitogenomics of *Hesperelaea*, an extinct genus of Oleaceae. *Gene* **2016**, *594*, 197–202. [[CrossRef](#)] [[PubMed](#)]
46. Zedane, L.; Hong-Wa, C.; Murienne, J.; Jeziorski, C.; Baldwin, B.G.; Besnard, G. Museumics illuminate the history of an extinct, paleoendemic plant lineage (*Hesperelaea*, Oleaceae) known from an 1875 collection from Guadalupe Island, Mexico. *Biol. J. Linn. Soc.* **2016**, *117*, 44–57. [[CrossRef](#)]
47. Straub, S.C.K.; Parks, M.; Weitemier, K.; Fishbein, M.; Cronn, R.C.; Liston, A. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *Am. J. Bot.* **2012**, *99*, 349–364. [[CrossRef](#)]
48. Berger, B.A.; Han, J.; Sessa, E.B.; Gardner, A.G.; Shepherd, K.A.; Ricigliano, V.A.; Jabaily, R.S.; Howarth, D.G. The unexpected depths of genome-skimming data: A case study examining Goodeniaceae floral symmetry genes. *Appl. Plant Sci.* **2017**, *5*, 1700042. [[CrossRef](#)]
49. Govindarajulu, R.; Parks, M.; Tennessen, J.A.; Liston, A.; Ashman, T.L. Comparison of nuclear, plastid, and mitochondrial phylogenies and the origin of wild octoploid strawberry species. *Am. J. Bot.* **2015**, *102*, 544–554. [[CrossRef](#)]
50. Sun, M.; Soltis, D.E.; Soltis, P.S.; Zhu, X.; Burleigh, J.G.; Chen, Z. Deep phylogenetic incongruence in the angiosperm clade Rosidae. *Mol. Phylogenet. Evol.* **2015**, *83*, 156–166. [[CrossRef](#)]
51. Liu, S.H.; Edwards, C.E.; Hoch, P.C.; Raven, P.H.; Barber, J.C. Genome skimming provides new insight into the relationships in *Ludwigia* section *Macrocarpon*, a polyploid complex. *Am. J. Bot.* **2018**, *105*, 875–887. [[CrossRef](#)]
52. Mathews, S.; Lavin, M.; Sharrock, R.A. Evolution of the phytochrome gene family and its utility for phylogenetic analyses of angiosperms. *Ann. Mo. Bot. Gard.* **1995**, *82*, 296–321. [[CrossRef](#)]
53. Mathews, S.; Tsai, R.C.; Kellogg, E.A. Phylogenetic structure in the grass family (Poaceae): Evidence from the nuclear gene phytochrome B. *Am. J. Bot.* **2000**, *87*, 96–107. [[CrossRef](#)] [[PubMed](#)]
54. WCSP. World Checklist of Selected Plant Families. Facilitated by the Royal Botanic Gardens, Kew. 2020. Published on the Internet. Available online: <http://wcsp.science.kew.org/> (accessed on 4 June 2020).
55. Banfi, E. *Chrysojasminum*, a new genus for *Jasminum* sect. *Alternifolia* (Oleaceae, Jasmineae). *Nat. Hist. Sci.* **2014**, *1*, 3–6. [[CrossRef](#)]
56. Bianconi, M.; Hackel, J.; Vorontsova, M.S.; Alberti, A.; Arthan, W.; Burke, S.V.; Duvall, M.R.; Kellogg, E.A.; Lavergne, S.; McKain, M.; et al. Continued adaptation of C₄ photosynthesis after an initial burst of changes in the Andropogoneae grasses. *Syst. Biol.* **2020**, *69*, 445–461. [[CrossRef](#)] [[PubMed](#)]
57. Kearse, M.; Moir, R.; Wilson, A.; Stones-Havas, S.; Cheung, M.; Sturrock, S.; Buxton, S.; Cooper, A.; Markowitz, S.; Duran, C.; et al. GENEIOUS Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **2012**, *28*, 1647–1649. [[CrossRef](#)]
58. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797. [[CrossRef](#)]
59. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Meth.* **2012**, *9*, 357–359. [[CrossRef](#)]
60. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.; Pham, S.; Pribelski, A.D.; et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **2012**, *19*, 455–477. [[CrossRef](#)]
61. Katoh, K.; Misawa, K.; Kuma, K.; Miyata, T. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **2002**, *30*, 3059–3066. [[CrossRef](#)]
62. Besnard, G.; Christin, P.A.; Malé, P.J.; Coissac, E.; Lhuillier, E.; Lauzeral, C.; Vorontsova, M.S. From museums to genomics: Old herbarium specimens shed light on a C₃ to C₄ transition. *J. Exp. Bot.* **2014**, *65*, 6711–6721. [[CrossRef](#)]
63. Besnard, G.; Bianconi, M.E.; Hackel, J.; Manzi, S.; Vorontsova, M.S.; Christin, P.A. Herbarium genomics retrace the origins of C₄-specific carbonic anhydrase in Andropogoneae (Poaceae). *Bot. Lett.* **2018**, *165*, 419–433. [[CrossRef](#)]
64. Patel, R.K.; Jain, M. NGS QC Toolkit: A toolkit for quality control of next generation sequencing data. *PLoS ONE* **2012**, *7*, e30619. [[CrossRef](#)] [[PubMed](#)]
65. Löytynoja, A. Phylogeny-aware alignment with PRANK. *Meth. Mol. Biol.* **2014**, *1079*, 155–170.
66. Minh, B.Q.; Schmidt, H.A.; Chernomor, O.; Schrempf, D.; Woodhams, M.D.; von Haeseler, A.; Lanfear, R. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **2020**, *37*, 1530–1534. [[CrossRef](#)] [[PubMed](#)]

67. Kalyaanamoorthy, S.; Minh, B.Q.; Wong, T.K.F.; von Haeseler, A.; Jermini, L.S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Meth.* **2017**, *14*, 587–589. [[CrossRef](#)]
68. Hoang, D.T.; Chernomor, O.; von Haeseler, A.; Minh, B.Q.; Vinh, L.S. UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **2018**, *35*, 518–522. [[CrossRef](#)] [[PubMed](#)]
69. Lanfear, R.; Frandsen, P.B.; Wright, A.M.; Senfeld, T.; Calcott, B. PartitionFinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol. Biol. Evol.* **2017**, *34*, 772–773. [[CrossRef](#)]
70. Phillips, M.J.; Delsuc, F.; Penny, D. Genome-scale phylogeny and the detection of systematic biases. *Mol. Biol. Evol.* **2004**, *21*, 1455–1458. [[CrossRef](#)]
71. Minh, B.Q.; Hahn, M.W.; Lanfear, R. New methods to calculate concordance factors for phylogenomic datasets. *Mol. Biol. Evol.* **2020**, *37*, 2727–2733. [[CrossRef](#)]
72. Kiew, R.; Baas, P. *Nyctanthes* is a member of Oleaceae. *Proc. Ind. Acad. Sci.* **1984**, *93*, 349–358.
73. George, K.; Geethamma, S. Cytological and other evidences for the taxonomic position of *Nyctanthes arbor-tristis*. *Curr. Sci.* **1984**, *53*, 439–441.
74. Huang, C.H.; Zhang, C.; Liu, M.; Hu, Y.; Gao, T.; Qi, J.; Ma, H. Multiple polyploidization events across Asteraceae with two nested events in the early history revealed by nuclear phylogenomics. *Mol. Biol. Evol.* **2016**, *33*, 2820–2835. [[CrossRef](#)] [[PubMed](#)]
75. Waterhouse, R.M.; Seppey, M.; Simão, F.A.; Manni, M.; Ioannidis, P.; Klioutchnikov, G.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **2018**, *35*, 543–548. [[CrossRef](#)] [[PubMed](#)]
76. Zhang, F.; Ding, Y.; Zhu, C.D.; Zhou, X.; Orr, M.C.; Scheu, S.; Luan, Y.X. Phylogenomics from low-coverage whole-genome sequencing. *Meth. Ecol. Evol.* **2019**, *10*, 507–517. [[CrossRef](#)]
77. Smith, S.A.; Donoghue, M.J. Rates of molecular evolution are linked to life history in flowering plants. *Science* **2008**, *322*, 86–89. [[CrossRef](#)]
78. Larracuente, A.M.; Sackton, T.B.; Greenberg, A.J.; Wong, A.; Singh, N.D.; Sturgill, D.; Zhang, Y.; Oliver, B.; Clark, A.G. Evolution of protein-coding genes in *Drosophila*. *Trends Genet.* **2008**, *24*, 114–123. [[CrossRef](#)]
79. Yang, L.; Gaut, B.S. Factors that contribute to variation in evolutionary rate among *Arabidopsis* genes. *Mol. Biol. Evol.* **2011**, *28*, 2359–2369. [[CrossRef](#)]

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Appendix for Chapter 1

Supporting information includes the following items:

Table S1. List of Oleaceae accessions analyzed in our study, with their taxonomy, accession number and origin.

Table S2. List of species used as outgroups in our phylogenetic analyses.

Figure S1. Full representation of the midpoint-rooted maximum likelihood phylogenetic tree of the *phy* gene family in Oleaceae.

Figure S2. Maximum likelihood topology of Oleaceae estimated from the partitioned analysis of the four datasets with corresponding concordance factors of nodes.

Figure S3. Maximum likelihood phylogenetic tree of Oleaceae based on the non-transformed nrDNA cluster alignment.

Table S1. List of Oleaceae accessions analyzed in our study, with their taxonomy, accession number and origin (including country and collection). *N = naturalized plant in the introduced range; BG = Botanical Garden; CEFE = Centre d'Ecologie Fonctionnelle et Evolutive, Montpellier; Herbarium acronyms follow Thiers (2019); ^h Sample directly removed from the herbarium specimen. ⁿ Newly analyzed accession.

Tribe (Subtribe)	Species	Accession no/cultivar name	Country (Collection)
Myxopyreae	<i>Dimetra craibiana</i> Kerr	S. Suddee et al. 2014 (K) ⁿ	Thailand
	<i>Myxopyrum nervosum</i> Blume	S.C. Chin 13831 [P04255167] ^{h,n}	Malaysia
	<i>Myxopyrum smilacifolium</i> (Wall.) Blume	J.E. Vidal 6056 [P03424472] ^{h,n}	Vietnam
	<i>Nyctanthes arbor-tristis</i> L.	C. Parma 9713 (M) ^{h,n}	India
Fontanesieae	<i>Fontanesia fortunei</i> Carrière	T. Joßberger 1829 (BONN)	China (Bonn BG)
Forsythieae	<i>Abeliophyllum distichum</i> Nakai	M.W. Chase 3881 (K)	Korea (Kew BG)
	<i>Forsythia</i> × <i>mandschurica</i> Uyeki	18-40-1968 (JFRU21)	China (Montreal BG)
Jasmineae	<i>Chrysojasminum fruticans</i> (L.) Banfi	G. Besnard 01-2016 (P) ⁿ	France (CEFE)
	<i>Jasminum didymum</i> G.Forst.	L. Barrabe 1312 [P01062395] ^{h,n}	New Caledonia
	<i>Jasminum lanceolaria</i> Roxb.	L. Jin-Kui 794 [P04254370] ^{h,n}	China
	<i>Jasminum nummularifolium</i> Baker	J.N. Labat 3712 [P00527351] ⁿ	Comoro
	<i>Jasminum pauciflorum</i> Benth.	D. Bilivogui 193 [P00853672] ^{h,n}	Guinea
	<i>Menodora integrifolia</i> (Cham. & Schltdl.) Steud.	A.L. Cabrera 29169 [P04492682] ^{h,n}	Argentina
Oleeae (Fraxininae)	<i>Fraxinus americana</i> L.	968-79 (R.E. Weaver & J. Nickerson)	USA (Arnold Arboretum)
	<i>Fraxinus angustifolia</i> Vahl	G. Besnard 42-2014 (P)	France
	<i>Fraxinus ornus</i> L.	Cefe-B2 (P. Saumitou-Laprade)	France (CEFE)
	<i>Fraxinus quadrangulata</i> Michx.	14654 (G.W. Letterman)	USA (Arnold Arboretum)
	<i>Fraxinus xanthoxyloides</i> (G.Don) Wall. ex A.DC.	Kepos 45350 (K)	Afghanistan (Kew BG)
Oleeae (Ligustrinae)	<i>Ligustrum lucidum</i> W.T.Aiton	G. Besnard 01-2019 (P) ⁿ	France (N*)
	<i>Ligustrum ovalifolium</i> Hassk.	G. Besnard 01-2018 (P) ⁿ	Portugal (CEFE, N*)
	<i>Syringa persica</i> L.	cv. "Laciniata" ⁿ	Unknown (cultivated)
	<i>Syringa pubescens</i> Turcz. subsp. <i>microphylla</i> (Diels) M.C.Chang & X.L.Chen	cv. "Superba" / G. Besnard 02-2015 (P)	China (Toulouse BG)
	<i>Syringa tomentella</i> Bureau & Franch. subsp. <i>yunnanensis</i> (Franch.) J.Y.Chen & D.Y.Hong	R. Lancaster 934	China (Arnold Arboretum)
	<i>Syringa vulgaris</i> L.	G. Besnard 02-1997 (P)	France
Oleeae (Schreberinae)	<i>Comoranthus obconicus</i> Knobl.	C. Mas 98 [P00176499] ^{h,n}	Mayotte
	<i>Schrebera swietenoides</i> Roxb.	W.S. Kurz 2312 (K) ^h	Myanmar
	<i>Schrebera trichoclada</i> Welw.	P.J. Greenway et al. 15382 (MO) ^h	Tanzania

Table S1, end.

Tribe (Subtribe)	Species	Accession no/cultivar name	Country (Collection)
Oleaceae (Oleinae)	<i>Cartrema americana</i> (L.) Raf.	CHW-OA (MO)	USA
	<i>Chengiodendron marginatus</i> (Champ. ex Benth.) C.B.Shang <i>et al.</i>	J.P.W. Woo 138 [P04074276] ^{h, n}	Hong Kong, China
	<i>Chionanthus</i> aff. <i>brassii</i> (Kobuski) Kiew	J.F. Molino 3078 (MPU)	Papua New Guinea
	<i>Chionanthus compactus</i> Sw.	C.M. Taylor 8582 [P03868414] ^{h, n}	Puerto Rico
	<i>Chionanthus ligustrinus</i> (Sw.) Pers.	A.H. Liogier 15164 [P03384280] ^h	Dominican Republic
	<i>Chionanthus parkinsonii</i> (Hutch.) Bennet & Raizada	M. Van de Bult 1222 (M) ^h	Thailand
	<i>Chionanthus</i> aff. <i>ramiflorus</i> Roxb.	M.E. Polane 24397 (K) ^h	Vietnam
	<i>Chionanthus retusus</i> Lindl. & Paxton	coll. KEW-13008	China (Kew BG)
	<i>Chionanthus rupicolus</i> (Lingelsh.) Kiew	W. Takeuchi et al. 15149 (K) ^h	Papua New Guinea
	<i>Chionanthus</i> aff. <i>thorelii</i> (Gagnep.) P.S.Green	M.F. Newman 2145 [P00577947]	Cambodia
	<i>Chionanthus</i> aff. <i>trichotomus</i> (Vell.) P.S.Green	St.G. Beck 25118 (M) ^h	Bolivia
	<i>Chionanthus virginicus</i> L.	coll. KEW-1976292	USA (Kew BG)
	<i>Forestiera angustifolia</i> Torr.	R. Kathy 1861, coll. 1997-0035-100	USA
	<i>Forestiera isabelae</i> Hammel & Cornejo	de Hammel & Perez 24248 (K) ^h	Costa Rica
	<i>Forestiera pubescens</i> Nutt. var. <i>pubescens</i>	LBJWC-0204	USA
	<i>Haenianthus salicifolius</i> Griseb.	A.H. Liogier 12182 [P04255168] ^{h, n}	Dominican Republic
	<i>Hesperelaea palmeri</i> A.Gray	E. Palmer 81 (MO) ^h	Guadalupe, Mexico
	<i>Nestegis cunninghamii</i> (Hook.f.) L.A.S.Johnson	Kepos 45352 (K), coll. KEW-1966-67114	New Zealand
	<i>Nestegis sandwicensis</i> (A.Gray) O.Deg., I.Deg. & L.A.S.Johnson	T. Flynn 6329 (MPU)	Hawaii, USA
	<i>Noronhia emarginata</i> (Lam.) Poir.	T. Flynn 6331 (MPU)	Hawaii, USA (N*)
	<i>Noronhia lowryi</i> Hong-Wa	J. Razanatsoa 686-1 (TAN)	Madagascar
	<i>Noronhia nilotica</i> (Oliv.) Hong-Wa & Besnard	White 886 (MO) ⁿ	Gabon
	<i>Noronhia peglerae</i> (C.H. Wright) Hong-Wa & Besnard	O. Maurin 1766 (PET)	South Africa
	<i>Notelaea longifolia</i> Vent.	L.A. Craven et al. 10154 [P04255187] ^h	Australia
	<i>Olea europaea</i> L. subsp. <i>laperrinei</i> (Batt. & Trab.) Cif.	coll. CEFE-Adjelella 10	Algeria (CEFE)
	<i>Olea javanica</i> (Blume) Knobl.	G. Besnard s.n. (MPU)	Indonesia
	<i>Olea lancea</i> Lam.	G. Besnard s.n. (P)	Reunion
	<i>Olea paniculata</i> R.Br.	C. Lambrides 1 (MPU)	Australia
	<i>Osmanthus armatus</i> Diels.	G. Besnard 02-2013 (P)	China (Toulouse BG)
	<i>Osmanthus austrocaledonicus</i> (Vieill.) Knobl.	J.K. Munzinger 1662 [P00354333]	New Caledonia
	<i>Osmanthus decorus</i> (Boiss. & Balansa) Kasapliligil	Accession Z-1	Caucasus (KEITH arboretum)
	<i>Phillyrea angustifolia</i> L.	Restinclières	France
	<i>Picconia excelsa</i> (Aiton) DC.	PT-0-BR-20110182-45	Madeira, Portugal
	<i>Priogymnanthus hasslerianus</i> (Chodat) P.S.Green	T. Rojas 10694 [P03384284] ^{h, n}	Paraguay

Table S2. List of species used as outgroups in our phylogenetic analyses. Number of genome project, accession nos and DNA regions are indicated.

Family	Species	Genome no (Plant isolate)	Accession no (DNA region)
Acanthaceae	<i>Avicennia marina</i> (Forssk.) Vierh.	PRJNA629068 (RG18007)	CM0231175 (<i>phyB</i>), CM0231176 (<i>phyB</i>), CM023179 (<i>ΨphyE</i>)
		PRJNA628464	NC_047414.1 (plastome)
		PRJNA6290689	SRR11912464 (mtDNA/nrDNA)
Pedaliaceae	<i>Sesamum indicum</i> L.	PRJNA268358 (Zhongzhi no 13)	XM_011073075 (<i>phyB</i>), XM_011102453 (<i>phyB</i>), XM_011086986 (<i>phyE</i>), SRR1055197 (mtDNA/nrDNA)
		PRJNA78529	NC_016433.2 (plastome)
Solanaceae	<i>Capsicum annuum</i> L.	PRJNA186921	CA05g16200 (<i>phyB</i>), CA01g16010 (<i>phyB</i>), CA02g12340 (<i>phyE</i>), JX270811.1 9 (plastome), KJ865410.1 (mtDNA), SRR653499 (nrDNA)
Solanaceae	<i>Solanum lycopersicum</i> L.	PRJNA119	Solyc05g053410 (<i>phyB</i>), Solyc01g059870 (<i>phyB</i>), Solyc02g071260 (<i>phyE</i>), DQ347959.1 (plastome), MF034192.1 (mtDNA), NW_020442480.1 (nrDNA)

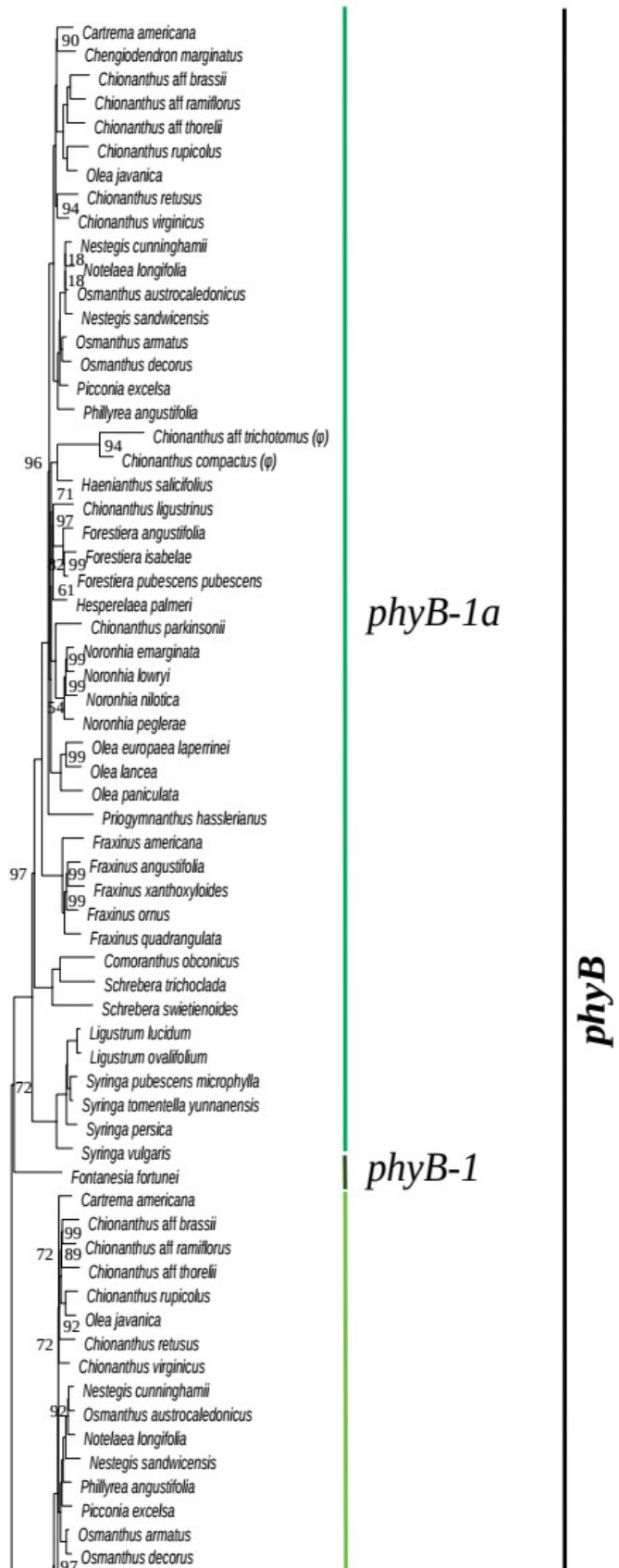


Figure S1. Full representation of the midpoint-rooted maximum likelihood phylogenetic tree of the *phy* gene family in Oleaceae. Only ultrafast bootstrap values inferior to 100 are indicated on nodes.

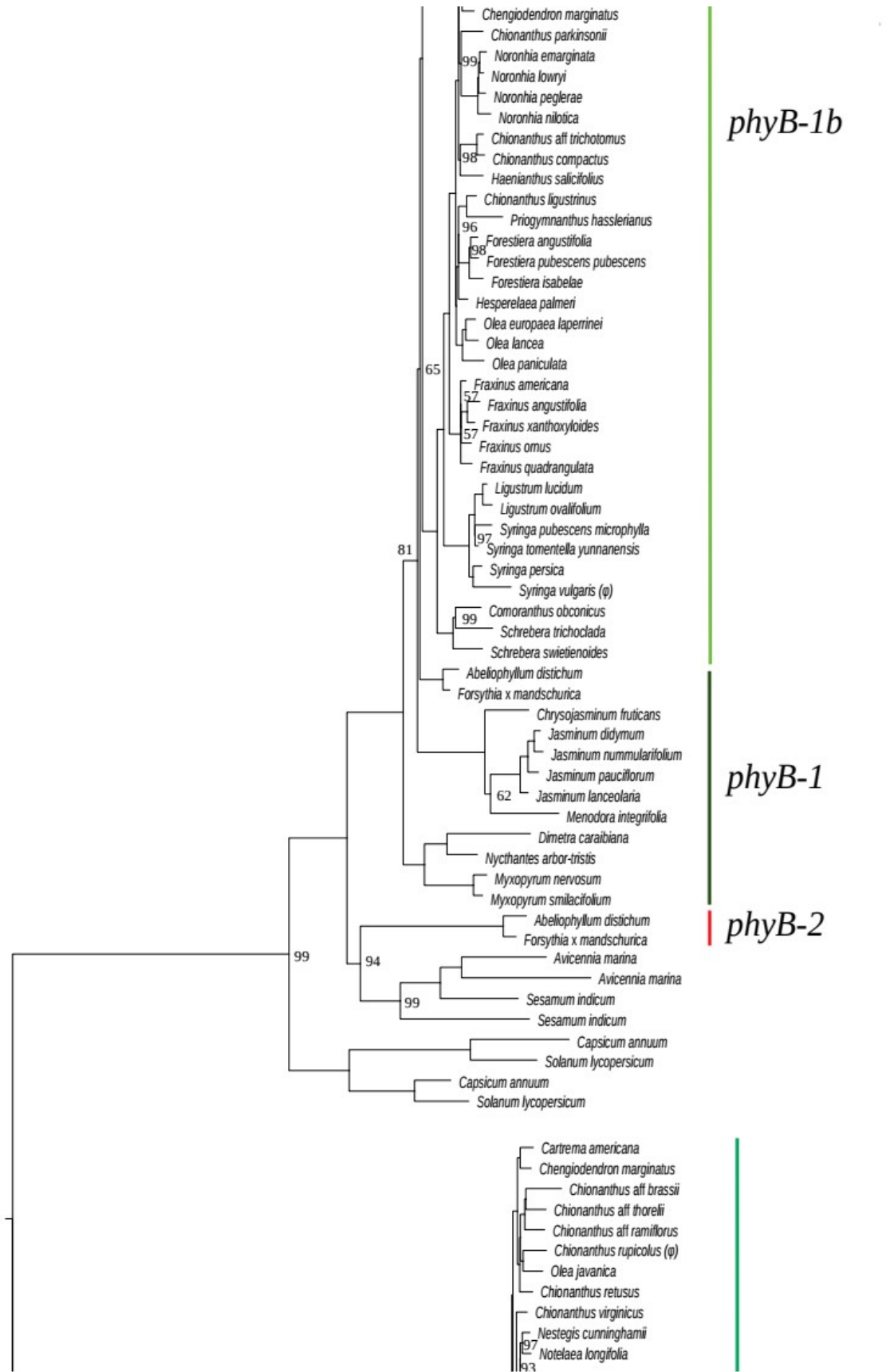


Figure S1, continued.

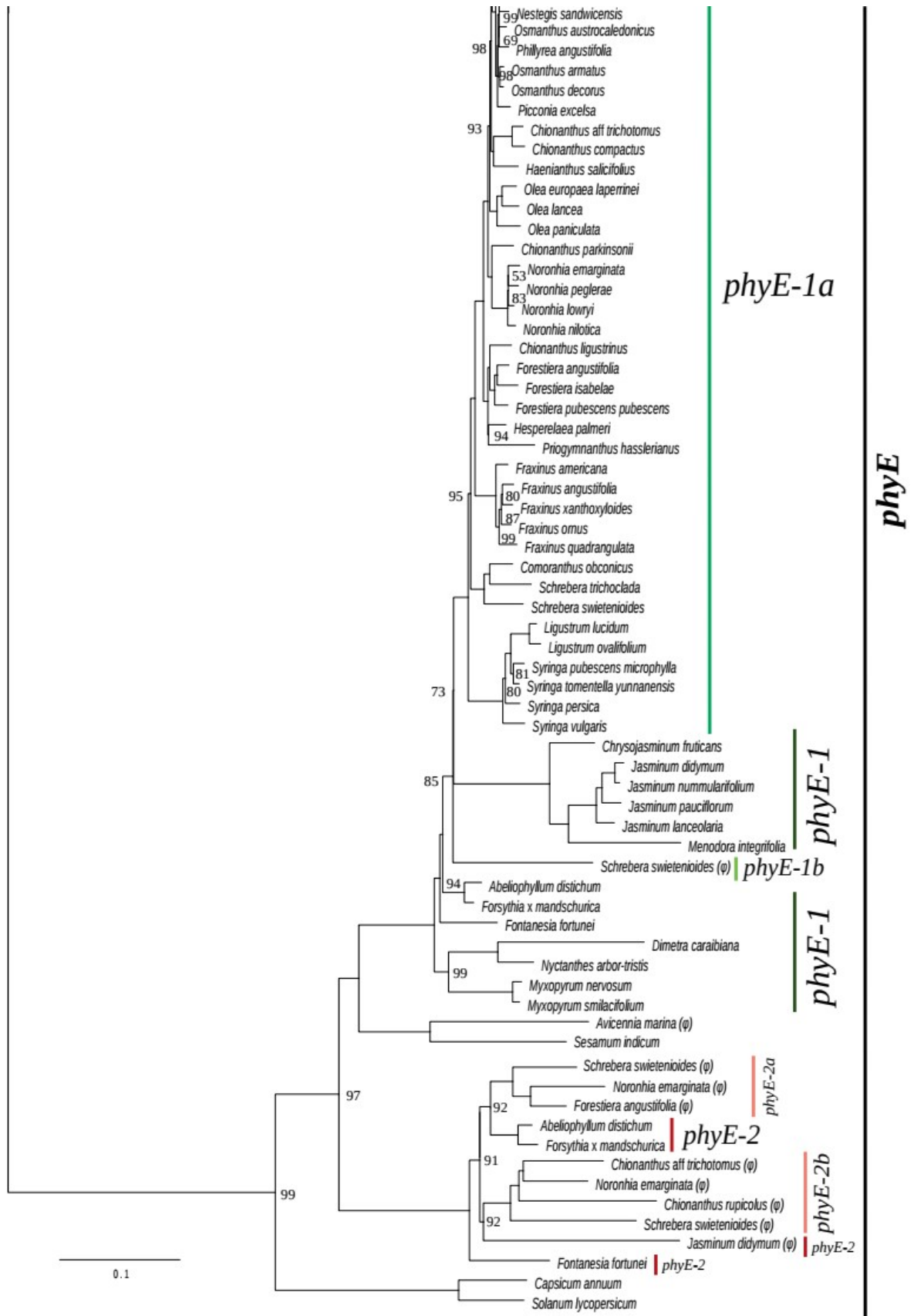
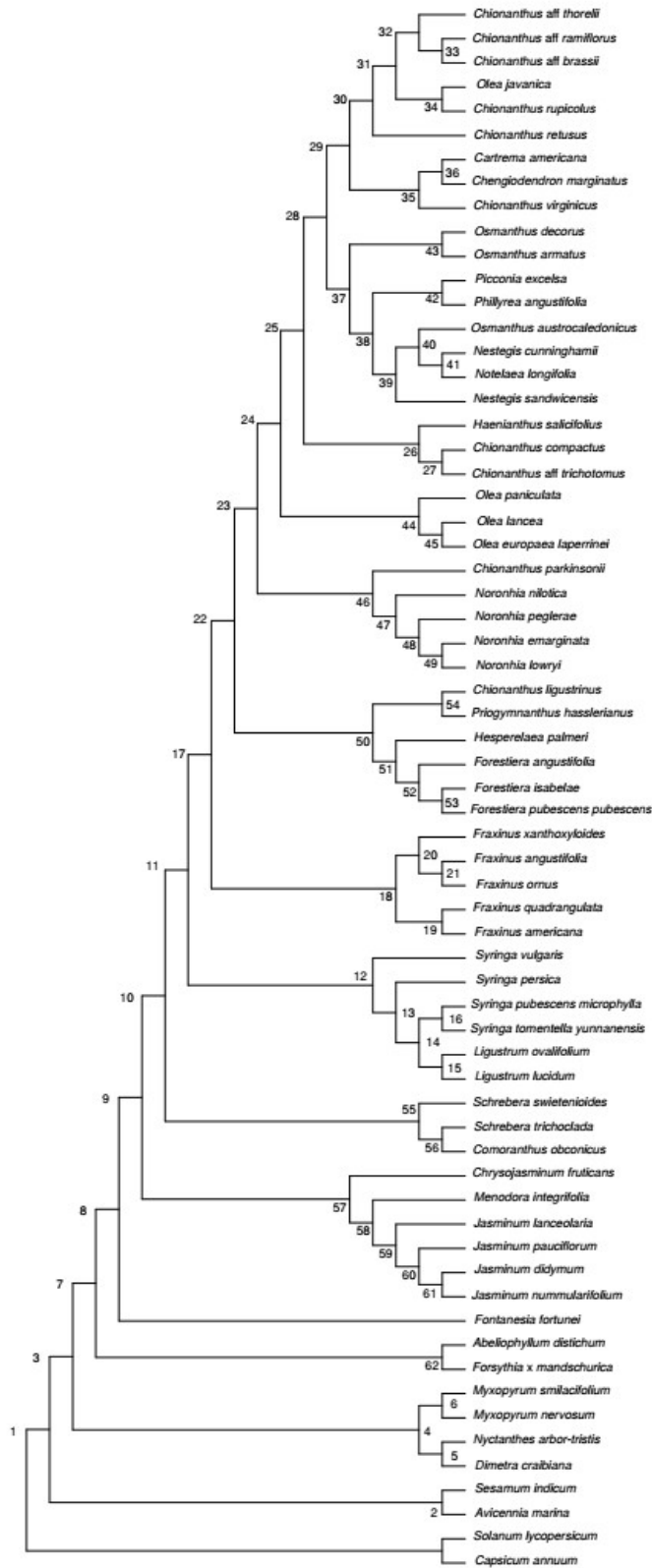


Figure S1, end.



Node	gCF	sCF
1	100	95.6
2	100	71.4
3	100	67.9
4	75	57.7
5	100	75.9
6	100	92.3
7	50	51.9
8	25	29.5
9	100	35.0
10	100	68.4
11	75	35.3
12	100	78.3
13	50	50.6
14	100	72.2
15	100	91.9
16	75	44.2
17	100	59.5
18	100	80.5
19	75	57.5
20	100	77.2
21	25	30.7
22	100	66.5
23	25	37.4
24	25	37.3
25	75	37.1
26	75	46.9
27	100	89.1
28	100	60.0
29	50	47.3
30	50	46.5
31	75	50.0
32	50	40.8
33	50	44.4
34	100	73.2
35	25	43.2
36	75	50.4
37	100	72.9
38	50	52.5
39	50	60.2
40	50	61.0
41	50	31.4
42	25	46.6
43	75	62.1
44	100	71.4
45	100	64.1
46	75	54.6
47	100	89.5
48	50	55.1
49	75	52.8
50	75	44.4
51	25	40.5
52	100	80.8
53	25	40.8
54	50	47.0
55	100	61.7
56	100	50.1
57	100	92.6
58	75	38.8
59	100	86.7
60	100	69.8
61	75	75.8
62	100	86.1

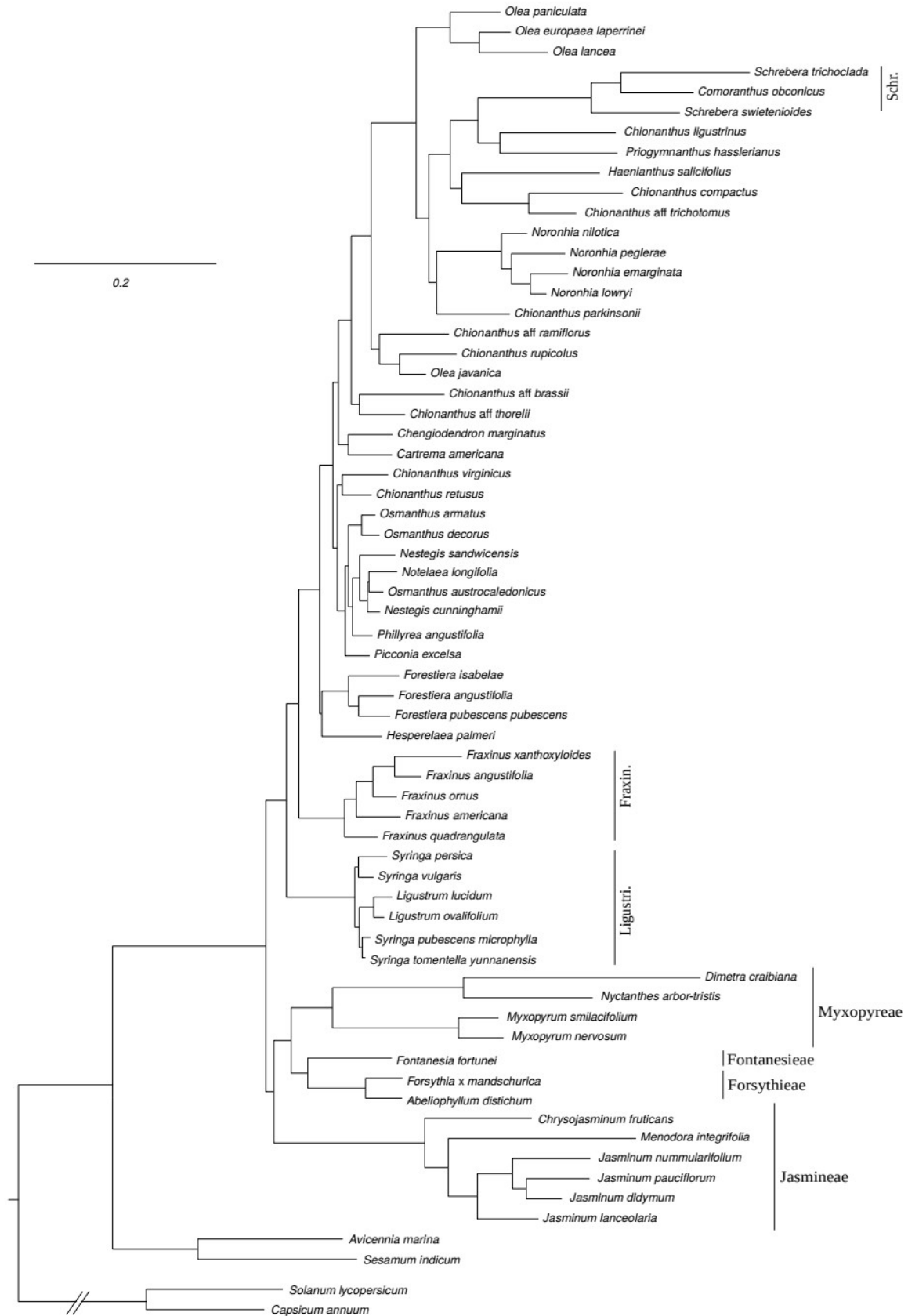


Figure S3. Maximum likelihood phylogenetic tree of Oleaceae based on the non-transformed nrDNA cluster alignment. The tree was rooted on the split with Solanaceae.

Chapter 2

Positive selection on non-photosynthetic genes drives parallel acceleration of plastid genome evolution in two Oleaceae lineages

Raimondeau P., Van de Paer C., Phillippe H., Christin P.-A.*, Besnard G.*

*Co-senior authors

Abstract

Rates of molecular evolution vary dramatically across genes and genomes. The multiple causes of this heterogeneity are difficult to disentangle. Organellar genomes of photosynthetic plants are thought to evolve under strong purifying selection and usually show little variation in evolutionary rates. Accelerated plastid evolution has been evidenced in several lineages, but the influence of selection has been generally disregarded. Here, we analyze the rates of plastome evolution among an extensive sampling of 117 Oleaceae. Plastid-specific acceleration of evolutionary rates was observed in two Oleaceae lineages. Comparison of the rates of synonymous and non-synonymous substitutions between the plastid and mitochondrial genomes revealed that this acceleration is mostly driven by non-synonymous changes. Using codon models, we showed that these patterns are driven by positive selection on a subset of non-photosynthetic genes. These shifts coincide with transitions to non-strict maternal inheritance of plastids, and we hypothesize that potential biparental transmission generates intracellular competition between undifferentiated plastids in zygotes, with strong selection for mutations that increase plastid proliferation leading to an excess of non-synonymous mutations. Our analyses show that selective processes can explain the acceleration of plastid evolution in some plant lineages.

Key-words: biparental inheritance, chloroplast genome, intracellular competition, Oleaceae, positive selection.

Introduction

Rates of molecular evolution can vary considerably. This variation is found at different scales, and can be genome-wide or locus-specific, between distinct lineages or within a given species (Nabholz *et al.*, 2008; Yang & Gaut, 2011; Smith & Keeling, 2015). Determining the factors driving this variation is a major goal in molecular evolution and has been the focus of many studies (Wolfe *et al.*, 1989a; Martin & Palumbi, 1993; Woolfit & Bromham, 2005; Bromham, 2009, 2011). The rate of DNA sequence evolution is affected by two factors: the rate at which mutations arise and the rate at which they get fixed. Features affecting either of these two rates will influence the overall evolutionary rates. Many factors have been suggested to have an influence, including environmental energy, metabolic rate, generation time, population size, recombination rate, genome copying processes, and natural selection (Bousquet *et al.*, 1992; Laroche *et al.*, 1997; Barraclough & Savolainen, 2001; Davies *et al.*, 2004; Wright *et al.*, 2006; Smith & Donoghue, 2008; Woolfit, 2009; Gillman *et al.*, 2010; Lynch, 2010; Lanfear *et al.*, 2013; Bromham *et al.*, 2015).

Disentangling the forces acting on evolutionary rates can be difficult, so that the drivers of differences among specific lineages are often unknown.

In plants, genetic information is shared between the nucleus and two organellar genomes located in plastids (plastome, also often referred to as chloroplast genome) and mitochondria (mitogenome). Organelle genomes encode functions essential for cell survival and are usually maternally inherited (Birky, 1995). Each cell contains many organelles, inside which organellar genomes can be present in multiple copies (Bendich, 1987). High-copy number, usual uniparental inheritance and the consecutive absence of sexual recombination are hallmarks of organellar genome evolution and make them an interesting system to investigate variation in rates of evolution as they can tell a different story from the recombining biparentally inherited nuclear genome. Among photosynthetic plants, variation in organelle evolutionary rates is generally limited and substitution rates of plastid protein-coding genes are usually lower than those of nuclear coding genes, but higher than in the mitogenome (Wolfe *et al.*, 1987, 1989b; Drouin *et al.*, 2008; Smith & Keeling, 2015). The importance of organelles in cellular metabolism and their extreme genomic sequence conservation support the idea that both organellar genomes predominantly evolve under strong purifying selection (Bock *et al.*, 2014). However, evidence for accelerated evolution of organellar genomes has been reported in several groups of photosynthetic plants, and can be restricted to particular loci (Guisinger *et al.*, 2008; Barnard-Kubow *et al.*, 2014), one (Guisinger *et al.*, 2010; Schwarz *et al.*, 2017; Shrestha *et al.*, 2019) or both organelles (Sloan *et al.*, 2012) or represent a global increase across all genomic compartments (Smith & Donoghue, 2008). A correlation between an acceleration of substitution rates and structural changes (i.e. inversions and gene or intron losses) has also been reported in several lineages (Jansen *et al.*, 2007; Barnard-Kubow *et al.*, 2014; Weng *et al.*, 2014; Zhu *et al.*, 2016; Schwarz *et al.*, 2017). These structural changes appear to be themselves correlated with the accumulation of repetitive DNA (Lee *et al.*, 2007; Haberle *et al.*, 2008; Guisinger *et al.*, 2011; Weng *et al.*, 2014). Such combination of increased substitution rates and structural changes has been explained by a disruption of the DNA repair/recombination/replication machinery (Guisinger *et al.*, 2008; Guisinger *et al.*, 2011; Weng *et al.*, 2014). The discussion of variation in evolutionary rates of organellar genomes has therefore been mainly centered on mutation processes. With the exception of cyto-nuclear coevolution (Sloan *et al.*, 2012; Williams *et al.*, 2019), selection has mainly been seen as being affected by the accelerated rates rather than as a potential driving force. Differential substitution rates between functional groups of genes, reported in several groups (Guisinger *et al.*, 2008; Sloan *et al.*, 2014; ; Park *et al.*, 2017), suggest that selection plays a more fundamental driving role. Genome-wide study of plastid rates in *Silene* and several Geraniaceae (Guisinger 2008, Sloan 2012; Sloan 2014; Park

2017) indeed reported elevated dN/dS ratio in a partly overlapping subset of genes. Positive selection was invoked however the potential selective agents were not discussed. Evidence for positive selection is limited outside of these groups as most study were restricted to a handful of genes or one taxa (Barnard)Passiflora Campanulacea +mitochondria

The olive tree family (Oleaceae) contains about 700 species, encompassing long-lived trees, shrubs and even one herb (Green, 2004). Previous evidence suggests different rates of evolution in some lineages, either affecting nuclear genes and/or one or both organellar genomes (Dupin *et al.*, 2020). Indeed, two lineages in particular (i.e. tribe Jasmineae and a monophyletic lineage of subtribe Ligustrinae, referred to as “Core Ligustrinae”) are characterized by markedly longer branches compared to sister clades in phylogenetic trees based on plastomes, but not necessarily in those based on mitochondrial or nuclear genes (Dupin *et al.*, 2020; Dong *et al.*, 2022). In one of these (Jasmineae), important reorganization of the plastomes and variation in gene and repeat content were also reported (Lee *et al.*, 2007). The plastomes of some groups of Oleaceae therefore seem characterized by increased rates of substitutions and structural changes, although the drivers of these patterns remain to be assessed. In particular, it is unclear whether these longer branches represent a plastid-specific phenomenon, or arise from lineage-specific factors (e.g. generation time, population size), in which case all genomic compartments would be affected. Rigorous comparison of the rates of evolution between these clades and other Oleaceae are needed to untangle the respective roles of mutation and selection processes in accelerated evolution.

In this study, we assess the importance of selective pressures in changes of evolutionary rates of organellar genomes. We first assemble plastomes and mitochondrial genes for a large panel of Oleaceae species, quantify rearrangements and infer phylogenomic trees to (i) verify that increased rates of substitutions and rearrangements characterize only some Oleaceae lineages. Independently for synonymous substitutions, which mainly reflect neutral processes, and non-synonymous substitutions, which are affected by selection, we then compare the substitution rates between the mitochondrial and plastid genes to (ii) determine whether both organellar genomes and (iii) both types of substitutions are simultaneously affected. We finally use codon models to (iv) test for selective shifts on specific organellar protein-coding genes driving non-synonymous substitution rates. Our investigations reveal a previously unreported effect of selective shifts on organellar genome evolutionary rates, potentially linked to differential selection regimes on non-photosynthetic genes following non-strict maternal transmission in some lineages.

Material and methods

Taxon sampling, DNA sequencing and organellar genomic sequence assembly

Our sampling consisted of 117 species representing all tribes and subtribes of Oleaceae as defined by Wallander & Albert (2000), plus two outgroup species (*Avicennia marina* – Acanthaceae and *Sesamum indicum* – Pedaliaceae), for a total of 119 species in the dataset. It included species from the two groups previously identified as having longer branches in the plastome phylogeny (Dupin *et al.*, 2020): eight Core Ligustrinae (*Ligustrum* spp., *Syringa pubescens* subsp. *microphylla*, and *S. tomentella* subsp. *yunnanensis*), and nine Jasmineae (representing the three genera of the tribe: *Chrysojasminum*, *Jasminum*, and *Menodora*).

New sequencing data were generated for five species (Supporting Information Table S4): three *Jasminum* (tribe Jasmineae: *J. adenophyllum*, *J. mackeeorum*, and *J. spectabile*) and two *Noronhia* (tribe *Oleeae*, subtribe *Oleinae*: *N. brevituba* and *N. candicans*). These additional samples were either collected from the field or from herbarium collections and sequenced with the approach described in Dupin *et al.*, (2020). Sequencing data were already available for the remaining 114 species, for which complete plastomes were already assembled (Van de Paer *et al.*, 2018; Olofsson *et al.*, 2019; Dupin *et al.*, 2020; Salmona *et al.*, 2020; Yu *et al.*, 2020). When available, complete mitogenomes or sequences for 37 mitochondrial protein-coding genes were also retrieved from previous studies (Supporting Information Table S4; Van de Paer *et al.*, 2016, 2018; Olofsson *et al.*, 2019; Dupin *et al.*, 2020). Previously unassembled organellar genomic sequences, including those of newly sequenced species or mitochondrial genes for a few species, were generated using the genome-walking method described in Dupin *et al.*, (2020). Plastomes were annotated using PGA (Qu *et al.*, 2019). The annotation of tRNA genes in all plastomes was further verified with tRNAscan-SE v2.0.7 (Chan & Lowe, 2019).

Structure, organization and content of plastomes

The 119 plastomes were aligned using progressiveMauve (Darling *et al.*, 2004) with default parameters to define collinear blocks and identify potential rearrangements. Where multiple rearrangements were identified, we used GRIMM (Tesler, 2002) to infer the minimum number of rearrangements based on these collinear blocks. Repeats over > 30 bp were identified using BLAST of each plastome against itself, after removing one inverted repeat copy (-word_size 16 -evalue 0.01; Altschul *et al.*, 1990), retaining only the best hit (according to e-value) among overlapping hits using the package GenomicRanges in R (Lawrence *et al.*, 2013). Organization maps were generated with the package GenoPlotR (Guy *et al.*, 2010).

Phylogenetic analyses

Protein-coding sequences were extracted and each gene was individually aligned as codons using the package DECIPHER v2 (Wright, 2016) in R. Genes from each organellar genome were concatenated, and a maximum-likelihood (ML) phylogeny was generated for each organelle using IQ-Tree v2.0.5 (Minh *et al.*, 2020). We used an edge-linked proportional partition model and assessed branch support with 1000 ultrafast bootstrap replicates (Chernomor *et al.*, 2016; Hoang *et al.*, 2018). The best partition scheme (genes, codon positions, or genes and codon positions) for each dataset was determined with PartitionFinder v2.1.1 (Lanfear *et al.*, 2017), and the best-fit evolutionary model for each partition was selected according to the best BIC score with ModelFinder (Kalyaanamoorthy *et al.*, 2017) as implemented in IQ-Tree.

Evolutionary rate estimations

The two organellar genomes produced compatible topologies (Supporting Information Figure S1), but the one based on the plastome has better statistical support. To facilitate rate comparisons, all follow-up analyses were performed with the topology obtained on plastomes.

Changes in branch lengths are the result of three main parameters: elapsed time, mutation rate, and selective pressures. To untangle the last two, synonymous substitution rates (mainly neutral, and thus approximating the basal mutation rate) and non-synonymous substitution rates (changes in the encoded amino acid) were estimated separately using the free-ratio model implemented in the program codeml (PAML4; Yang, 2007). In this model, the expected number of non-synonymous substitutions per site (d_N) and the expected number of synonymous substitutions per site (d_S) are free to vary among branches. For these analyses, all protein-coding sequences from each organellar genome were concatenated to produce rates per genome. Gaps and missing data were excluded from the alignment (cleandata = 1). The two organellar genomes were compared using the set of d_S values estimated for all branches. If some species experienced lineage-specific factors that accelerated their evolution or if branches correspond to a longer divergence time, the d_S of the two organellar genomes are expected to increase proportionally, so that a linear relationship is expected. Conversely, if one of the two organellar genomes accumulates higher relative amounts of synonymous substitutions in one lineage, a different slope is expected for values assigned to branches belonging to this group. To test whether relative rates of evolution between the two organellar genomes differ between the two groups characterized by long branches in previous studies (i.e. Core Ligustrinae and Jasmineae) and between the rest of Oleaceae, we used a linear model, where the slope of the relationship between the d_S of plastid coding sequence and the d_S of mitochondrial coding sequence differed among the three sets of branches (interaction with the

group identity). Significance of each model coefficient was assessed with an ANOVA. Similar analyses were performed using the d_N values estimated for each branch. We also optimized the free-ratio model on each single gene to determine whether particular genes showed distinct profiles compared to the whole coding compartment. Plastid genes were grouped into two functional categories: genes related to photosynthetic functions (photosystems, NADH, cytochromes) and non-photosynthetic genes (RNA polymerase, ribosomal proteins, transcription factors, and other unknown functions; Bock *et al.*, 2007). These groups were used as categorical variables in the analyses.

Selective pressure estimations

Codon models were used to specifically test, separately for the plastome and the mitogenome and then for each gene within these genomes, whether selective pressures shifted in the set of branches assigned to one of the two clades with longer branches in the plastome phylogenetic tree (i.e. Core Ligustrinae and Jasmineae). Two different codon models implemented in codeml were optimized and statistically compared. In the null model, all branches evolve under the same d_N/d_S ratio. In the alternative model, *a priori* selected branches (“foreground branches”) are allowed to have a different d_N/d_S ratio from the other branches (“background branches”). Branches belonging to Core Ligustrinae and Jasmineae (including the stem branches) were selected as two sets of “foreground branches”, with a distinct d_N/d_S ratio (three ratios). The two models were compared using a likelihood-ratio test with the appropriate degrees of freedom and a Bonferroni correction was applied to account for multiple testing. For the concatenated datasets, we also compared a model with both Core Ligustrinae and Jasmineae considered as a single set of foreground branches (two ratios).

Results

Rearrangements and changes in plastome content are restricted to Jasmineae and Core Ligustrinae

As in most angiosperms, Oleaceae plastomes exhibit the canonical quadripartite structure of two single copy regions (i.e. large single copy and small single copy regions) separated by a large inverted repeat (Bock, 2007). Inverted repeat boundaries are remarkably stable across the family and important shifts (encompassing complete genes) were only observed in Jasmineae and Core Ligustrinae (Supporting Information Figure S2). Inversions events were only detected in Jasmineae, with one event shared by all investigated Jasmineae and few species-specific rearrangements, with the most drastic rearrangements observed in a small shrub of this tribe, *Menodora integrifolia* (Supporting Information Figure S3). Repeats longer than 30 bp accounted for a significantly larger

genome proportion in both Jasmineae and Core Ligustrinae compared to other Oleaceae (Supporting Information Figure S4; ANOVA; p -value < 0.05), reaching 17% in *M. integrifolia*. In Jasmineae and Core Ligustrinae, the median number of repeats longer than 50 bp was 16, as opposed to six in other Oleaceae. Larger repeats (>150 bp) are restricted to these two clades. The plastome of *M. integrifolia* is the only one to contain repeated fragments larger than 1000 bp (up to 5000 bp).

Gene content is quite conserved across the family and only few gene losses were detected, all in Jasmineae (Supporting Information Table S1). Firstly, *clpP*, which encodes the plastid subunit of a protease complex essential in plastid protein turnover, was not found in *M. integrifolia*. The two introns of *clpP* have also been lost in *Jasminum* species, and exon sequences are highly divergent in this genus. In the Jasmineae *Chrysojasminum fruticans*, the coding sequence of *clpP* shows divergence levels similar to that of other Jasmineae, but *clpP* introns are still present and are even more conserved than exons (raw pairwise distances with *Olea europaea* are 0.30 and 0.036 in exons and introns, respectively). Exonic divergence is also significantly greater than intronic divergence in Core Ligustrinae (t -test; p -value < 0.001), a pattern not observed in other Oleaceae (Supporting Information Table S2). Secondly, *accD*, which encodes the plastid subunit of acetyl-CoA carboxylase (which catalyses the first and rate-limiting step of lipid biosynthesis), has also been lost in Jasmineae. Lastly, variation in the number of copies of tRNA genes was also restricted to Jasmineae and *Ligustrum ovalifolium* (Core Ligustrinae; Supporting Information Table S1).

Rate increases are specific to the plastome and mainly driven by non-synonymous substitutions

Phylogenetic trees built on the complete set of plastid and mitochondrial protein-coding sequences for 117 taxa representative of all Oleaceae subtribes confirm increased root-to-tip distances in both Jasmineae and Core Ligustrinae for plastid sequences, but such increases are less apparent in the mitochondrial phylogeny (Supporting Information Figure S1). Phylogenetic trees built on synonymous substitutions revealed clearly longer branches in Jasmineae but not in Core Ligustrinae, while both groups exhibit markedly longer branches in trees based on non-synonymous substitutions (Figure 1). The regression slope of plastid d_s in relation to mitochondrial d_s is significantly different (ANOVA; p -value < 0.01) in Jasmineae (but not in Ligustrinae; p -value = 0.9) compared to other Oleaceae, indicating synonymous substitutions are accumulating faster in the plastomes of Jasmineae (Figure 2; adjusted $R^2 = 0.80$). These patterns suggest an increase in the underlying mutation rate of plastome in the whole Jasmineae group.

Regarding d_N , significant differences (p -value < 0.01) between the slopes of d_N were observed in Jasmineae and in Core Ligustrinae compared to other Oleaceae (Figure 2; adjusted $R^2 =$

0.88). In both groups, the portions of plastome corresponding to coding genes accumulate more non-synonymous substitutions than in other Oleaceae (Figure 2). Codon models supported three different d_N/d_S ratios in the plastomes of each of Jasmineae, Core Ligustrinae and the other Oleaceae (likelihood-ratio test, p -value < 0.05). The optimized d_N/d_S values over concatenated coding-sequences were higher in both Jasmineae ($\omega_J = 0.52$) and Core Ligustrinae ($\omega_{CL} = 0.96$) than in other Oleaceae ($\omega_B = 0.25$), indicating that a shift of selective pressures characterizes the two groups with increased plastid branch lengths.

Overall, our results indicate that the increased accumulation of substitutions happened specifically and independently in the plastomes of Jasmineae and Core Ligustrinae and was driven by an accelerated accumulation of both synonymous and non-synonymous substitutions in Jasmineae, but mostly by an excess of non-synonymous substitutions in Core Ligustrinae.

Non-photosynthetic genes undergo an excess of non-synonymous substitutions in Jasmineae and Core Ligustrinae

Individual codon models were used to determine which specific genes have an increased d_N/d_S in Jasmineae and/or Core Ligustrinae. For the mitogenome, we tested 37 protein-coding genes and did not identify any significant change in d_N/d_S in these groups. In plastomes, we detected one gene (*rbcL*) with a significant decrease in d_N/d_S , in both Jasmineae and Core Ligustrinae. Most changes, however, corresponded to increases of d_N/d_S , and among the 80 plastid genes investigated, 18 had a significant increase in d_N/d_S in branches belonging to Jasmineae and/or Core Ligustrinae (Figure 3). Most of the impacted genes encode proteins not involved in photosynthetic processes (15/18), and photosynthetic genes are statistically under-represented among those with an increased d_N/d_S (Fisher's exact test; p -value < 0.001). Out of 47 genes related to photosynthesis, an increased d_N/d_S was detected only in *atpA*, *atpB* and *ycf4* (Figure 3). Among the 15 non-photosynthetic genes with significant increases, d_N/d_S is greater than 1 in both Jasmineae and Core Ligustrinae for seven genes, whereas seven genes have a $d_N/d_S > 1$ in Jasmineae only, and one in Core Ligustrinae only (Figure 3). The d_N/d_S were especially high in Core Ligustrinae, with five genes reaching values above 3 (up to 5.7 for *infA*; Figure 3; Supporting Information Figure S5). Genes with a d_N/d_S ratio greater than 1 in one or both sets of foreground branches were reanalyzed constraining the d_N/d_S to a value of 1 in that set and the likelihood of these constrained models were compared to that of the unconstrained ones. After correcting for multiple testing, the d_N/d_S was significantly greater than 1 in *clpP* and *ycf1* in Jasmineae, and in *ycf1*, *ycf2* and *ycf4* in Core Ligustrinae.

To investigate further the observed changes in substitution rates at the gene level, d_S and d_N values per gene were compared between Jasmineae, Core Ligustrinae and other Oleaceae. Rates of

synonymous substitutions vary tremendously among genes, but the d_s per gene vary proportionally between each of Jasmineae and Core Ligustrinae and the rest of the family, with similar values in Core Ligustrinae, but on average a six time higher rate in Jasmineae (Figure 4a, c). The relationship between the synonymous rate in the other Oleaceae and the synonymous rate in either Jasmineae or Core Ligustrinae does not differ significantly between photosynthetic and non-photosynthetic genes (p -values = 0.07 and 0.6, respectively).

The patterns are strikingly different when comparing the d_N per gene, with similar trends in Jasmineae and Core Ligustrinae. For photosynthetic genes, the d_N vary proportionally between each of the two groups and the other Oleaceae, with ratios roughly equivalent to d_s estimates (3.76 vs 6.04 for Jasmineae and 1.72 vs 1.04 for Core Ligustrinae; Figure 4). These results indicate that non-synonymous substitutions on photosynthetic genes accumulate at a rate similar to synonymous substitutions. The relationship between the d_N of each of Jasmineae and Core Ligustrinae and other Oleaceae, however, differs significantly between photosynthetic and non-photosynthetic genes (p -values < 0.01 and < 0.05, respectively). In Jasmineae, the d_N of non-photosynthetic genes are about ten times larger than in other Oleaceae (or more than twice the values of photosynthetic genes), while the d_N of non-photosynthetic genes are about seven times higher in Core Ligustrinae than in other Oleaceae (or about four times higher than for photosynthetic genes; Figure 4b, d). These results show that all genes are characterized by more synonymous and non-synonymous substitutions in Jasmineae, but non-photosynthetic genes from the plastids of both Jasmineae and Core Ligustrinae underwent more than twice as much non-synonymous substitutions as expected based on their amount of synonymous substitutions.

Discussion

Selection drives evolutionary rate acceleration in plastomes of two Oleaceae lineages

Longer branches were previously reported in the plastid phylogeny in Core Ligustrinae and Jasmineae (Dupin *et al.*, 2020; Dong *et al.*, 2022). Our analyses confirmed that the relative increases in substitution rates in these two groups are restricted to the plastome (Supporting Information Figure S1) and thus do not result from species-specific factors that would also affect other genomic compartments. Jasmineae plastomes show higher numbers of rearrangements (Supporting Information Figure S3), as well as an increase in synonymous substitution rates compared to other Oleaceae (Figure 1), which affects all genes similarly (Figure 4a). Mutation processes have been invoked to explain accelerated global plastome evolution (Guisinger *et al.*, 2008; Guisinger *et al.*, 2011; Weng *et al.*, 2014). Faulty repair mechanisms can impact substitution rates and induce genomic rearrangements, if they fail to prevent illegitimate recombination.

However, most proteins involved in repair mechanisms are encoded by the nuclear genome and target both plastid and mitochondrial DNA (Carrie & Small, 2013). The pattern of divergence observed here would thus require the repair machinery to operate incorrectly only on plastomes.

An increase in non-synonymous substitution rates is observed in both Jasmineae and Core Ligustrinae (Figure 2) and non-photosynthetic genes are particularly impacted (Figure 4b, d). Some of these genes (in particular *accD* and *clpP*) are recurrently reported as fast-evolving in several groups of plants and have also been lost in several lineages with accelerated evolutionary rates (Guisinger *et al.*, 2008, 2010; Sloan *et al.*, 2012; Rockenbach *et al.*, 2016; Williams *et al.*, 2019). In Jasmineae, both *accD* and *clpP* or its introns are missing in some taxa (Supporting Information Table S1). While this could be interpreted as a sign of relaxation of selective pressures, exon sequences of *clpP* are more divergent than introns in both Jasmineae and Core Ligustrinae (when still present; Supporting Information Table S2). This surprising pattern was also reported in other lineages (Erixon & Oxelman, 2008; Barnard-Kubow *et al.*, 2014) and could be consistent with positive selection. Moreover, relaxed purifying selection generally cannot drive d_N/d_S well above 1, in contrast to what we observe for *clpP* and a subset of genes (Figure 3). Positive selection was further significantly supported compared to relaxed selection in several genes, using a conservative test. Therefore, positive selection seems to explain at least some of the excess of non-synonymous substitutions in Jasmineae and Core Ligustrinae.

Biparental inheritance and intracellular competition might drive change in selective pressures

Changes in plastome evolutionary rates are restricted to two clades of Oleaceae (Figure 1). We do not see lineage-specific features susceptible to impact plastome evolution in these two sets of species compared to other Oleaceae except for the mode of plastid inheritance. Indeed, characters allowing biparental inheritance have been identified precisely in these two groups of Oleaceae (Corriveau & Coleman, 1988; Zhang *et al.*, 2003; Liu *et al.*, 2004; Supporting Information Table S3). We thus hypothesize a link between biparental inheritance of plastids and their accelerated evolution. This association has previously been suggested based on the overlapping distribution of fast-evolving lineages and biparental inheritance (Ruhlman & Jansen, 2014), but the proposed mechanism involved recombination between parental copies in heteroplasmic cells. As plastids rarely fuse, evidence for plastome recombination is scarce (reviewed in Greiner *et al.*, 2015). Even in cases where biparental zygotes are produced, no recombination has been observed between parental organellar genomes (Birky, 1995; Matsushima *et al.*, 2008).

We propose another explanation for the coincidence in Jasmineae and Core Ligustrinae of putative biparental inheritance and elevated non-synonymous substitutions occurring specifically in

non-photosynthetic genes. Plant germline and reproductive tissues are constantly renewed from meristems that contain undifferentiated cells as well as plastid precursors, called proplastids (Bock, 2007; Lanfear, 2018). These immature plastids do not perform photosynthesis and remain undifferentiated in reproductive cells (Waters & Langdale, 2009). Biparental transmission will result in divergent proplastids sharing the same cell (i.e. heteroplasmy). Vegetative segregation usually leads to the stochastic loss of one haplotype as individuals grow (Birky, 2001). Yet, heteroplasmy provides an opportunity for intracellular selection if the two haplotypes confer different replication speeds for example (Chiu *et al.*, 1988). The most abundant type will be numerically advantaged during vegetative sorting, meaning that genes with mutations boosting proplastid multiplication are more likely to spread. Before they differentiate into chloroplasts, the probability of transmission of proplastids is thus directly determined by their capacity to propagate and not by the fitness of the individual that carries them. Once turned into chloroplasts, natural selection will act at the individual level and carriers of deleterious haplotypes will be eliminated from the population. Interestingly, it was recently demonstrated that the competitive ability of plastids in *Oenothera* was determined by fast-evolving repeat variation in the promoter region of *accD* and *ycf2* (Sobanski *et al.*, 2019), two genes that are also under positive selection in Core Ligustrinae (Figure 3). Within-cell variance is decreased by uniparental inheritance and by a strong transmission bottleneck during which the number of proplastids per cell is drastically reduced before their entrance into the germline (Birky, 1995). Biparental transmission will reduce the bottleneck efficiency and may thus promote the spread of selfish mutations promoting plastid competitiveness. Intracellular selection could explain the underrepresentation of photosynthetic genes in the fast-evolving set. However, heteroplasmy resulting from biparental inheritance can also promote selection at the intra-individual level and produce selective changes in plastome regardless of the functional category of the genes. More information regarding plastid abundance, cellular divisions and segregation dynamics would be necessary to untangle the respective forces of intracellular and intra-individual selection during plant development.

Conclusions

Rates of molecular evolution vary both among species and among genes within species, but the responsible mechanisms are not always well understood. In this work, we focus on the drivers of accelerated evolution of plastid genomes occurring in two lineages within the Oleaceae family. We show that this accelerated evolution is restricted to the plastid genome and does not concern the other organelle genome (mitogenome), ruling out organism-level processes. In one of the two lineages, the accelerated evolution is partially driven by a general increase of synonymous and non-

synonymous substitutions, but in all cases, the strongest increase is observed in non-synonymous substitutions, pointing to changes in selective pressures. Importantly, an increase of the ratio of non-synonymous to synonymous substitutions is mostly observed in non-photosynthetic genes in these two lineages, and this ratio often statistically exceeds one, pointing to positive selection. Because these shifts in selective pressures coincide with acquisitions of the ability to inherit plastids biparentally, we hypothesize that intracellular competition in the zygote offers an opportunity for selection on the plastid propagation rate. Such processes occurring in reproductive tissue before the proplastids differentiate into chloroplasts, only non-photosynthetic functions would be affected, explaining the observed patterns. We conclude that selective pressures, potentially linked to biparental inheritance, at least partially explain the parallel increase of rates of plastid evolution in two Oleaceae lineages. While other processes, such as defaulted DNA repair, might exacerbate the effects of selection, our investigation shows that minor modifications of the reproductive success can have major consequences for the long-term evolution of organellar genomes.

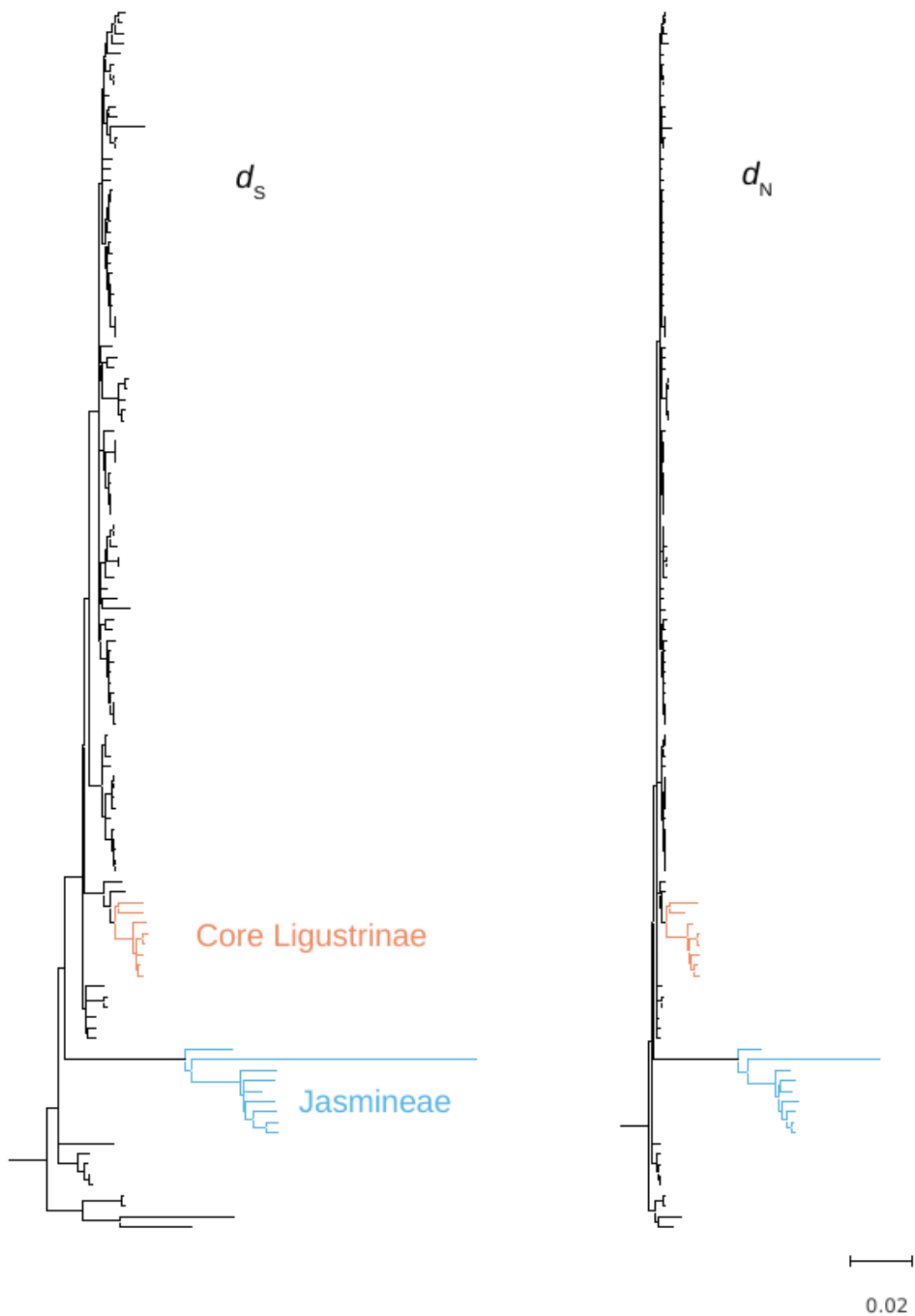


Figure 1. Phylogenetic trees of Oleaceae with branch lengths scaled to rate of synonymous (d_s) and non-synonymous substitutions (d_N) for plastid genomes. Rates were estimated in PAML under the free-ratio model using all protein-coding genes. The scale is in expected number of substitutions per site.

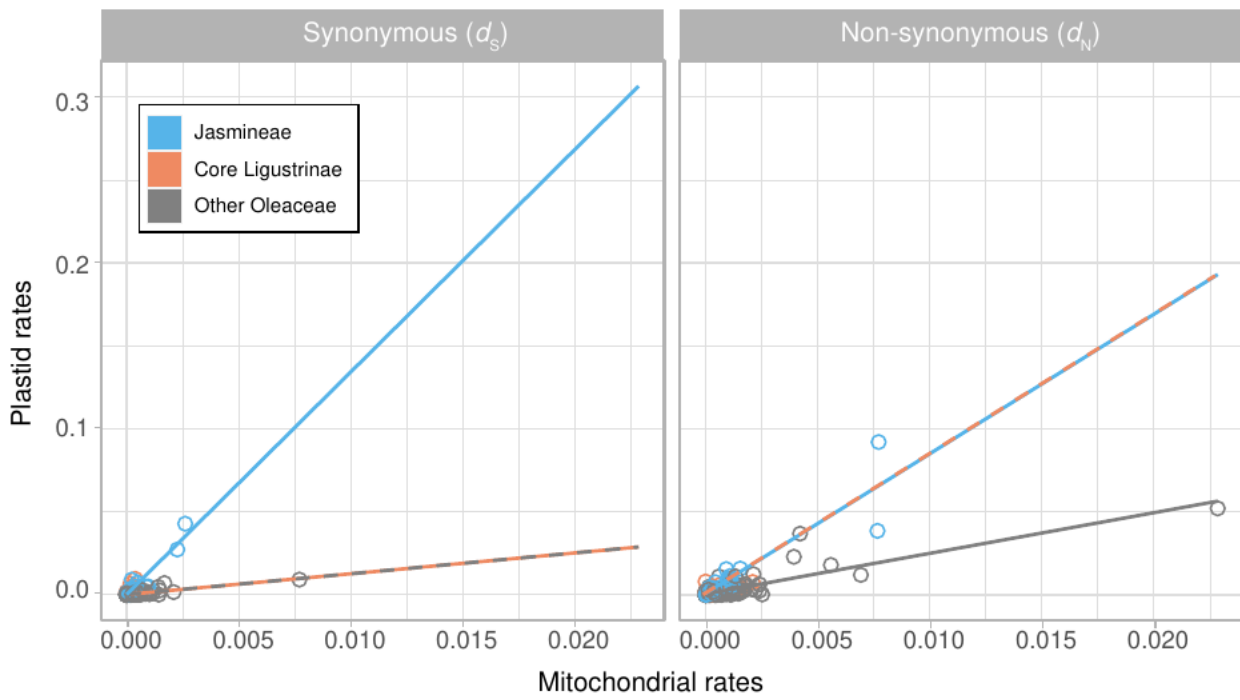


Figure 2. Relationship between plastome and mitogenome evolutionary rates in different subgroups of Oleaceae. Synonymous substitution rates are presented on the left panel, and non-synonymous rates are shown on the right. Rates are expressed in expected number of substitutions per site. Each point represents one branch and is colored according to the lineage it belongs to. Significantly different linear regression lines are shown.

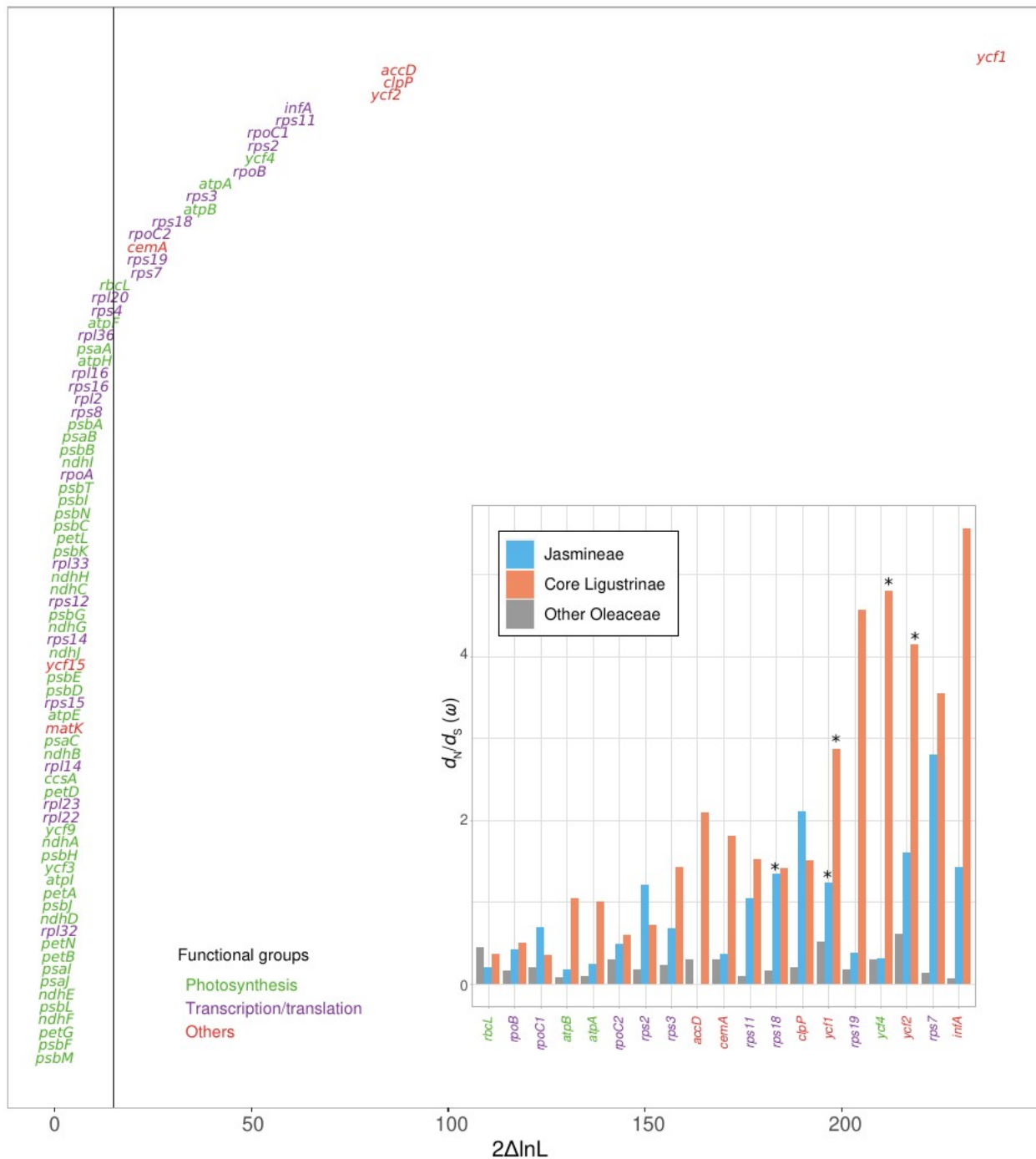


Figure 3. Alterations of selective pressures in plastid genes of Jasmineae and Core Ligustrinae. For the background plots, twice the difference of log-likelihood between the null model (single d_N/d_S for the whole tree) and the alternative model (different d_N/d_S in Jasmineae, Core Ligustrinae and the other Oleaceae) is indicated. The vertical black line indicates the significance threshold for a likelihood ratio test (from a χ^2 distribution with Bonferroni correction for multiple testing). Genes are ranked according to this value and colored based on their function. The insert shows d_N/d_S ratios (ω) in the three groups for the 19 genes with a significant change. Asterisks indicate genes where d_N/d_S is significantly greater than 1 (after correction).

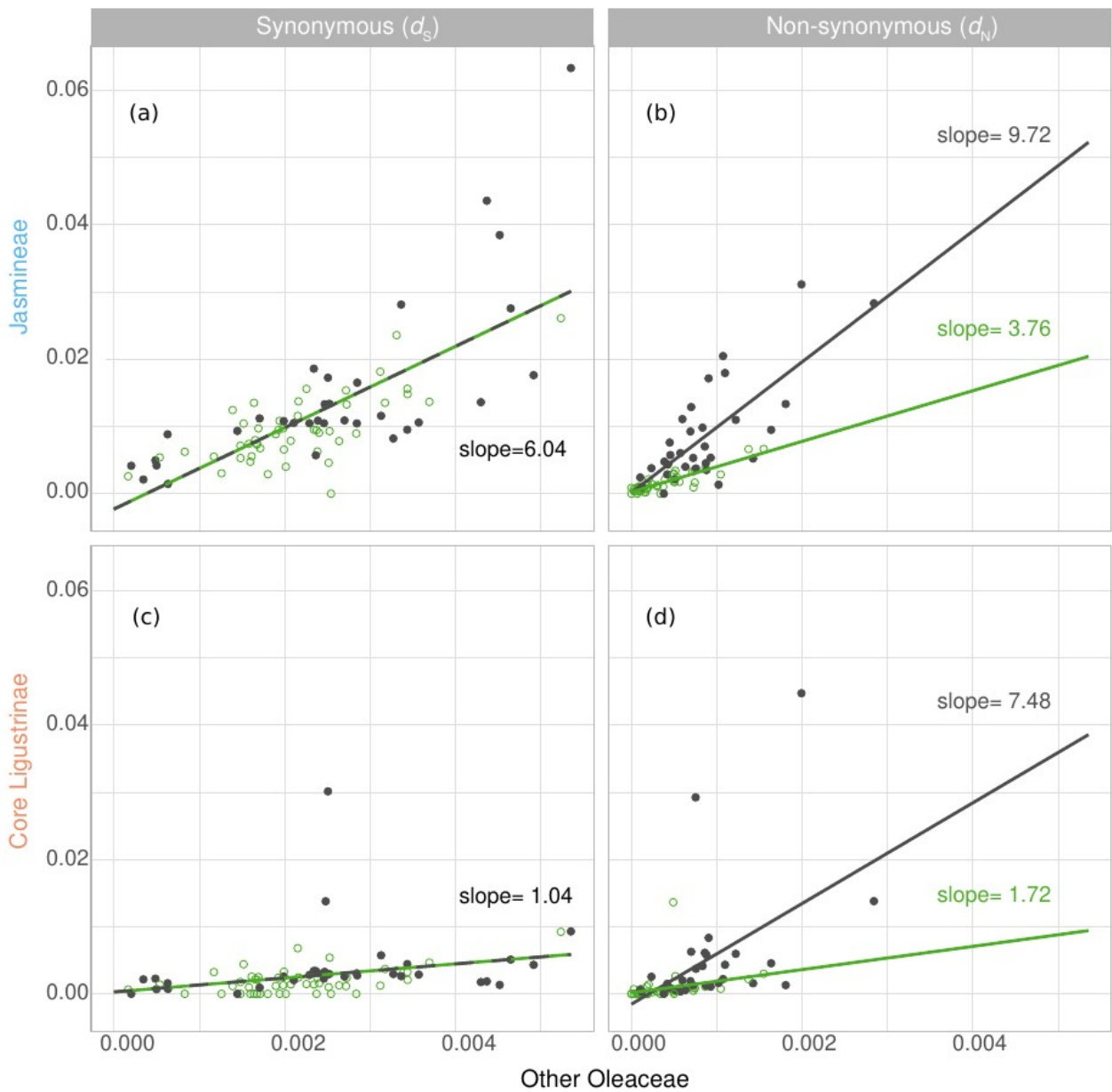


Figure 4. Relationship between plastid rates of synonymous (a, c) and non-synonymous substitutions (b, d) in Jasmineae (upper row) or Core Ligustrinae (lower row) and other Oleaceae. Each point represents a gene (mean rate value per group). Genes with functions related to photosynthesis are colored in green. Significantly different linear regression lines are shown.

Data availability

DNA sequence data supporting this study have been deposited at GenBank (NCBI). The accession numbers are listed in Supporting Information Table S4 (Supplementary Material online). Alignments and scripts are available on Github: https://github.com/praimondeau/plastome_evolution.

Author contributions

GB and PAC designed the study. GB and CVDP generated the data. PR analysed the data with input from GB, PAC and HP. PAC and PR wrote the paper with the help of GB. All authors commented on the final manuscript.

Acknowledgments

PR, CVDP and GB are members of the Laboratoire Evolution & Diversité Biologique (EDB), part of the LABEX TULIP managed by Agence Nationale de la Recherche (ANR; no. ANR-10-LABX-0041) and were funded by GeneRes (Occitanie-France Olive). We also acknowledge an Investissement d'Avenir grant of the ANR (CEBA: ANR-10-LABX-25-01). PAC is supported by a Royal Society Research Fellowship (grant number URF\R\180022). We thank Pierre-Marc Delaux, Jean Keller, Cyril Libourel and Camille Puginier for helpful discussions. We thank S. Manzi for lab assistance and the genotoul bioinformatics platform Toulouse Midi-Pyrenees for providing computing and storage resources. We are also grateful to all people who participated to plant sampling: H. Gryta (EDB), J.D. Thompson (CEFE), D. Stadie (Eisleben), J.M. Dosmann and K. Richardson (Arnold Arboretum, Harvard), E. Bellefroid (National Botanical Garden of Belgium), O. Maurin and D. Goyder (Royal Botanical Gardens, Kew), M. Gaudeul (Muséum National d'Histoire Naturelle de Paris), H. Esser (Munich Botanical Gardens), J. Razanatoa and F. Rakotonasolo (Parc de Tsimbazaza, Antananarivo), T. Josseberger and A. Krämer (Botanische Gärten der Universität Bonn), A. Rinfret-Pilo (Botanical Garden of Montreal), S. Blackwell (Botanical Garden of Phoenix), J. Munzinger (IRD Montpellier), R. Lima and G. Frey (Universidade de São Paulo), the herbarium of the Royal Botanical Garden of Edinburgh, The United States Department of Agriculture, and the Charles R. Keith Arboretum.

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215: 403–410.
- Barnard-Kubow KB, Sloan DB, Galloway LF. 2014. Correlation between sequence divergence and polymorphism reveals similar evolutionary mechanisms acting across multiple timescales in a rapidly evolving plastid genome. *BMC Evolutionary Biology* 14: 268.
- Barracough TG, Savolainen V. 2001. Evolutionary rates and species diversity in flowering plants. *Evolution* 55: 677–683.
- Bendich AJ. 1987. Why do chloroplasts and mitochondria contain so many copies of their genome? *BioEssays* 6: 279–282.
- Birky CW. 1995. Uniparental inheritance of mitochondrial and chloroplast genes: Mechanisms and evolution. *Proceedings of the National Academy of Sciences, USA* 92: 11331–11338.

- Birky CW. 2001. The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. *Annual Review of Genetics* 35: 125–148.
- Bock DG, Andrew RL, Rieseberg LH. 2014. On the adaptive value of cytoplasmic genomes in plants. *Molecular Ecology* 23: 4899–4911.
- Bock R. 2007. Structure, function, and inheritance of plastid genomes. In: Bock R, editor. *Cell and Molecular Biology of Plastids. Topics in Current Genetics*, vol. 19. Berlin (Heidelberg): Springer. p. 29–63.
- Bousquet J, Strauss SH, Doerksen AH, Price RA. 1992. Extensive variation in evolutionary rate of *rbcL* gene sequences among seed plants. *Proceedings of the National Academy of Sciences, USA* 89: 7844–7848.
- Bromham L. 2009. Why do species vary in their rate of molecular evolution? *Biology Letters* 5: 401–404.
- Bromham L. 2011. The genome as a life-history character: why rate of molecular evolution varies between mammal species. *Philosophical Transactions of the Royal Society B: Biological Sciences* 366: 2503–2513.
- Bromham L, Hua X, Lanfear R, Cowman PF. 2015. Exploring the relationships between mutation rates, life history, genome size, environment, and species richness in flowering plants. *American Naturalist* 185: 508–524.
- Carrie C, Small I. 2013. A reevaluation of dual-targeting of proteins to mitochondria and chloroplasts. *Biochimica et Biophysica Acta - Molecular Cell Research* 1833: 253–259.
- Chan PP, Lowe TM. 2019. tRNAscan-SE: searching for tRNA genes in genomic sequences. In: Kollmar M, editor. *Gene Prediction. Methods in Molecular Biology*, vol. 1962. New York (NY): Humana Press. p. 1–14.
- Chernomor O, von Haeseler A, Minh BQ. 2016. Terrace aware data structure for phylogenomic inference from supermatrices. *Systematic Biology* 65: 997–1008.
- Chiu WL, Stubbe W, Sears BB. 1988. Plastid inheritance in *Oenothera*: organelle genome modifies the extent of biparental plastid transmission. *Current Genetics* 13: 181–189.
- Corriveau JL, Coleman AW. 1988. Rapid screening method to detect potential biparental inheritance of plastid DNA and results for over 200 angiosperm species. *American Journal of Botany* 75: 1443–1458.
- Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research* 14: 1394–1403.
- Davies TJ, Savolainen V, Chase MW, Moat J, Barraclough TG. 2004. Environmental energy and evolutionary rates in flowering plants. *Proceedings of the Royal Society B: Biological Sciences* 271: 2195–2200.
- Drouin G, Daoud H, Xia J. 2008. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. *Molecular Phylogenetics and Evolution* 49: 827–831.
- Dong W, Li E, Liu Y, Xu C, Wang Y, Liu K, Cui X, Sun J, Suo Z, Zhang Z, *et al.* 2022. Phylogenomic approaches untangle early divergences and complex diversifications of the olive plant family. *BMC Biol.* 20:92.
- Dupin J, Raimondeau P, Hong-Wa C, Manzi S, Gaudeul M, Besnard G. 2020. Resolving the phylogeny of the olive family (Oleaceae): confronting information from organellar and nuclear genomes. *Genes* 11: 1508.

- Erixon P, Oxelman B. 2008. Whole-gene positive selection, elevated synonymous substitution rates, duplication, and indel evolution of the chloroplast *clpP1* gene. *PLoS One* 3: e1386.
- Gillman LN, Keeling DJ, Gardner RC, Wright SD. 2010. Faster evolution of highly conserved DNA in tropical plants. *Journal of Evolutionary Biology* 23: 1327–1330.
- Green PS. 2004. Oleaceae. In: Kadereit JW, editor. *Flowering Plants. Dicotyledons: Lamiales (except Acanthaceae including Avicenniaceae)*. Berlin (Heidelberg): Springer. p. 296–306.
- Greiner S, Sobanski J, Bock R. 2015. Why are most organelle genomes transmitted maternally? *BioEssays* 37: 80–94.
- Guisinger MM, Chumley TW, Kuehl JV, Boore JL, Jansen RK. 2010. Implications of the plastid genome sequence of *Typha* (Typhaceae, Poales) for understanding genome evolution in Poaceae. *Journal of Molecular Evolution* 70: 149–166.
- Guisinger MM, Kuehl JV, Boore JL, Jansen RK. 2008. Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. *Proceedings of the National Academy of Sciences, USA* 105: 18424–18429.
- Guisinger MM, Kuehl JV, Boore JL, Jansen RK. 2011. Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. *Molecular Biology and Evolution* 28: 583–600.
- Guy L, Kultima JR, Andersson SGE. 2010. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* 26: 2334–2335.
- Haberle RC, Fourcade HM, Boore JL, Jansen RK. 2008. Extensive rearrangements in the chloroplast genome of *Trachelium caeruleum* are associated with repeats and tRNA genes. *Journal of Molecular Evolution* 66: 350–361.
- Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution* 35: 518–522.
- Jansen RK, Cai Z, Raubeson LA, Daniell H, Depamphilis CW, Leebens-Mack J, Müller KF, Guisinger-Bellian M, Haberle RC, Hansen AK, *et al.* 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proceedings of the National Academy of Sciences, USA* 104: 19369–19374.
- Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* 14: 587–589.
- Lanfear R. 2018. Do plants have a segregated germline? *PLoS Biology* 16: e2005439.
- Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. 2017. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution* 34: 772–773.
- Lanfear R, Ho SYW, Jonathan Davies T, Moles AT, Aarssen L, Swenson NG, Warman L, Zanne AE, Allen AP. 2013. Taller plants have lower rates of molecular evolution. *Nature Communications* 4: 1879.
- Laroche J, Li P, Maggia L, Bousquet J. 1997. Molecular evolution of angiosperm mitochondrial introns and exons. *Proceedings of the National Academy of Sciences, USA* 94: 5722–5727.
- Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for computing and annotating genomic ranges. *PLoS Computational Biology* 9: e1003118.

- Lee HL, Jansen RK, Chumley TW, Kim KJ. 2007. Gene relocations within chloroplast genomes of *Jasminum* and *Menodora* (Oleaceae) are due to multiple, overlapping inversions. *Molecular Biology and Evolution* 24: 1161–1180.
- Liu Y, Cui H, Zhang Q, Sodmergen. 2004. Divergent potentials for cytoplasmic inheritance within the genus *Syringa*. A new trait associated with speciation. *Plant Physiology* 136: 2762–2770.
- Lynch M. 2010. Evolution of the mutation rate. *Trends in Genetics* 26: 345–352.
- Magee AM, Aspinall S, Rice DW, Cusack BP, Sémon M, Perry AS, Stefanović S, Milbourne D, Barth S, Palmer JD, *et al.* 2014. Localized hypermutation and associated gene losses in legume chloroplast genomes. *Genome Research* 24: 1052.
- Martin AP, Palumbi SR. 1993. Body size, metabolic rate, generation time, and the molecular clock. *Proceedings of the National Academy of Sciences, USA* 90: 4087–4091.
- Matsushima R, Hu Y, Toyoda K, Sodmergen, Sakamoto W. 2008. The model plant *Medicago truncatula* exhibits biparental plastid inheritance. *Plant Cell and Physiology* 49: 81–91.
- Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution* 37: 1530–1534.
- Nabholz B, Glémin S, Galtier N. 2008. Strong variations of mitochondrial mutation rate across mammals - The longevity hypothesis. *Molecular Biology and Evolution* 25: 120–130.
- Olofsson JK, Cantera I, Van de Paer C, Hong-Wa C, Zedane L, Dunning LT, Alberti A, Christin PA, Besnard G. 2019. Phylogenomics using low-depth whole genome sequencing: a case study with the olive tribe. *Molecular Ecology Resources* 19: 877–892.
- Park S, Ruhlman TA, Weng ML, Hajrah NH, Sabir JSM, Jansen RK. 2017. Contrasting patterns of nucleotide substitution rates provide insight into dynamic evolution of plastid and mitochondrial genomes of *Geranium*. *Genome Biology and Evolution* 9: 1766–1780.
- Qu XJ, Moore MJ, Li DZ, Yi TS. 2019. PGA: a software package for rapid, accurate, and flexible batch annotation of plastomes. *Plant Methods* 15: 50.
- Rockenbach K, Havird JC, Grey Monroe J, Triant DA, Taylor DR, Sloan DB. 2016. Positive selection in rapidly evolving plastid-nuclear enzyme complexes. *Genetics* 204: 1507–1522.
- Ruhlman TA, Jansen RK. 2014. The plastid genomes of flowering plants. In: Maliga P, editor. *Chloroplast Biotechnology: Methods in Molecular Biology (Methods and Protocols)*, vol. 1132. Totowa (NJ): Humana Press. p. 3–38.
- Ruhlman TA, Jansen RK. 2018. Aberration or analogy? The atypical plastomes of Geraniaceae. *Advances in Botanical Research* 85: 223–262.
- Salmona J, Olofsson JK, Hong-Wa C, Razanatsoa J, Rakotonasolo F, Ralimanana H, Randriamboavonjy T, Suescun U, Vorontsova MS, Besnard G. 2020. Late Miocene origin and recent population collapse of the Malagasy savanna olive tree (*Noronhia lowryi*). *Biological Journal of the Linnean Society* 129: 227–243.
- Schwarz EN, Ruhlman TA, Weng ML, Khiyami MA, Sabir JSM, Hajarrah NH, Alharbi NS, Rabah SO, Jansen RK. 2017. Plastome-wide nucleotide substitution rates reveal accelerated rates in Papilionoideae and correlations with genome features across legume subfamilies. *Journal of Molecular Evolution* 84: 187–203.
- Shrestha B, Weng ML, Theriot EC, Gilbert LE, Ruhlman TA, Krosnick SE, Jansen RK. 2019. Highly accelerated rates of genomic rearrangements and nucleotide substitutions in plastid

- genomes of *Passiflora* subgenus *Decaloba*. *Molecular Phylogenetics and Evolution* 138: 53–64.
- Sloan DB, Alverson AJ, Wu M, Palmer JD, Taylor DR. 2012. Recent acceleration of plastid sequence and structural evolution coincides with extreme mitochondrial divergence in the angiosperm genus *Silene*. *Genome Biology and Evolution* 4: 294–306.
- Sloan DB, Triant DA, Forrester NJ, Bergner LM, Wu M, Taylor DR. 2014. A recurring syndrome of accelerated plastid genome evolution in the angiosperm tribe Sileneae (Caryophyllaceae). *Molecular Phylogenetics and Evolution* 72: 82–89.
- Smith DR, Keeling PJ. 2015. Mitochondrial and plastid genome architecture: reoccurring themes, but significant differences at the extremes. *Proceedings of the National Academy of Sciences, USA* 112: 10177–10184.
- Smith SA, Donoghue MJ. 2008. Rates of molecular evolution are linked to life history in flowering plants. *Science* 322: 86–89.
- Sobanski J, Giavalisco P, Fischer A, Kreiner JM, Walther D, Schöttler MA, Pellizzer T, Golczyk H, Obata T, Bock R, *et al.* 2019. Chloroplast competition is controlled by lipid biosynthesis in evening primroses. *Proceedings of the National Academy of Sciences, USA* 116: 5665–5674.
- Sodmergen, Bai HH, He JX, Kuroiwa H, Kawano S, Kuroiwa T. 1998. Potential for biparental cytoplasmic inheritance in *Jasminum officinale* and *Jasminum nudiflorum*. *Sexual Plant Reproduction* 11: 107–112.
- Tesler G. 2002. GRIMM: genome rearrangements web server. *Bioinformatics* 18: 492–493.
- Van de Paer C, Bouchez O, Besnard G. 2018. Prospects on the evolutionary mitogenomics of plants: a case study on the olive family (Oleaceae). *Molecular Ecology Resources* 18: 407–423.
- Van de Paer C, Hong-Wa C, Jeziorski C, Besnard G. 2016. Mitogenomics of *Hesperelaea*, an extinct genus of Oleaceae. *Gene* 594: 197–202.
- Wallander E, Albert VA. 2000. Phylogeny and classification of Oleaceae based on *rps16* and *trnL-F* sequence data. *American Journal of Botany* 87: 1827–1841.
- Waters MT, Langdale JA. 2009. The making of a chloroplast. *The EMBO Journal* 28: 2861–2873.
- Weng ML, Blazier JC, Govindu M, Jansen RK. 2014. Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. *Molecular Biology and Evolution* 31: 645–659.
- Williams AM, Friso G, van Wijk KJ, Sloan DB. 2019. Extreme variation in rates of evolution in the plastid Clp protease complex. *The Plant Journal* 98: 243–259.
- Wolfe KH, Li WH, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences, USA* 84: 9054–9058.
- Wolfe KH, Sharp PM, Li WH. 1989a. Mutation rates differ among regions of the mammalian genome. *Nature* 337: 283–285.
- Wolfe KH, Sharp PM, Li WH. 1989b. Rates of synonymous substitution in plant nuclear genes. *Journal of Molecular Evolution* 29: 208–211.
- Woolfit M. 2009. Effective population size and the rate and pattern of nucleotide substitutions. *Biology Letters* 5: 417–420.

- Woolfit M, Bromham L. 2005. Population size and molecular evolution on islands. *Proceedings of the Royal Society B: Biological Sciences* 272: 2277–2282.
- Wright ES. 2016. Using DECIPHER v2.0 to analyze big biological sequence data in R. *The R Journal* 8: 352–359.
- Wright S, Keeling J, Gillman L. 2006. The road from Santa Rosalia: A faster tempo of evolution in tropical climates. *Proceedings of the National Academy of Sciences, USA* 103: 7718–7722.
- Yang L, Gaut BS. 2011. Factors that contribute to variation in evolutionary rate among *Arabidopsis* genes. *Molecular Biology and Evolution* 28: 2359–2369.
- Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* 24: 1586–1591.
- Yu X, Jiang W, Tan W, Zhang X, Tian X. 2020. Deciphering the organelle genomes and transcriptomes of a common ornamental plant *Ligustrum quihoui* reveals multiple fragments of transposable elements in the mitogenome. *International Journal of Biological Macromolecules* 165: 1988–1999.
- Zhang Q, Liu Y, Sodmergen. 2003. Examination of the Cytoplasmic DNA in Male Reproductive Cells to Determine the Potential for Cytoplasmic Inheritance in 295 Angiosperm Species. *Plant Cell Physiol.* 44:941–951.
- Zhu A, Guo W, Gupta S, Fan W, Mower JP. 2016. Evolutionary dynamics of the plastid inverted repeat: The effects of expansion, contraction, and loss on substitution rates. *New Phytologist* 209: 1747–1756.

Appendix for Chapter 2

Supporting information includes the following items:

Figure S1. Comparison between plastid and mitochondrial phylogenies

Figure S2. Physical maps summarizing inverted repeat boundary shifts detected in Oleaceae

Figure S3. Physical maps summarizing rearrangements detected in the long single copy (LSC) of Jasmineae plastomes

Figure S4. Distribution of repeat content of Oleaceae plastid genomes

Figure S5. Omega values for plastid genes in Jasmineae and Core Ligustrinae

Table S1. Gene and intron losses detected in Oleaceae

Table S2. cIPp exonic and intronic raw pairwise distances with *Olea europaea*

Table S3. List of Oleaceae species for which the inheritance mode of plastids has been reported

Table S4. List of Oleaceae accessions analyzed in our study and respective GenBank numbers

Supplementary notes. Investigating nuclear rates variations

References

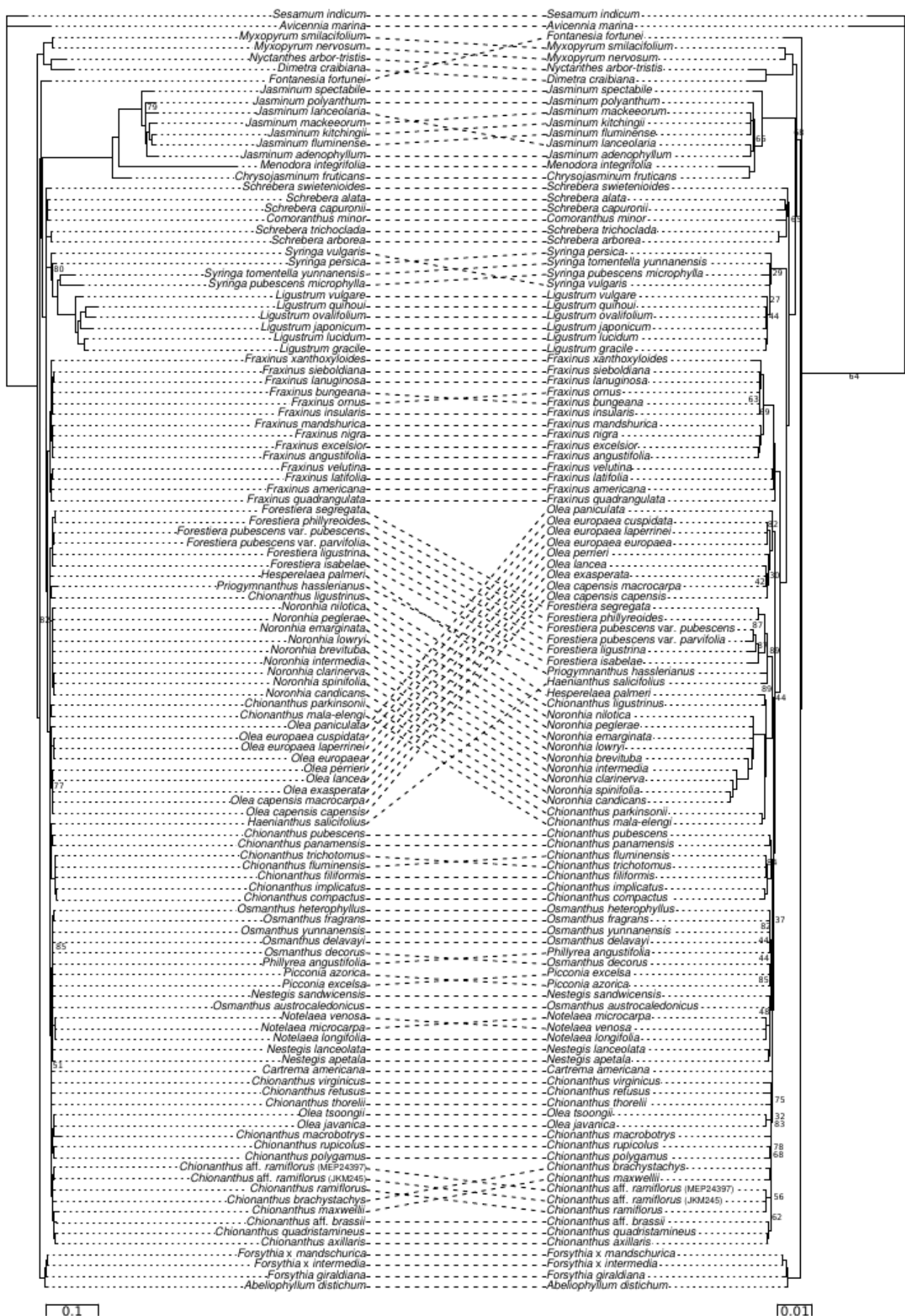


Figure S1. Comparison between plastid and mitochondrial topologies. Each tip of the plastid tree (left) is connected by a dashed line to the corresponding tip in the mitochondrial phylogram (right). Bootstraps values with support greater than 95 are masked. The scale is in substitution per site. Both topologies are largely congruent. Conflicts only arise at poorly-supported nodes in the mitochondrial phylogeny.

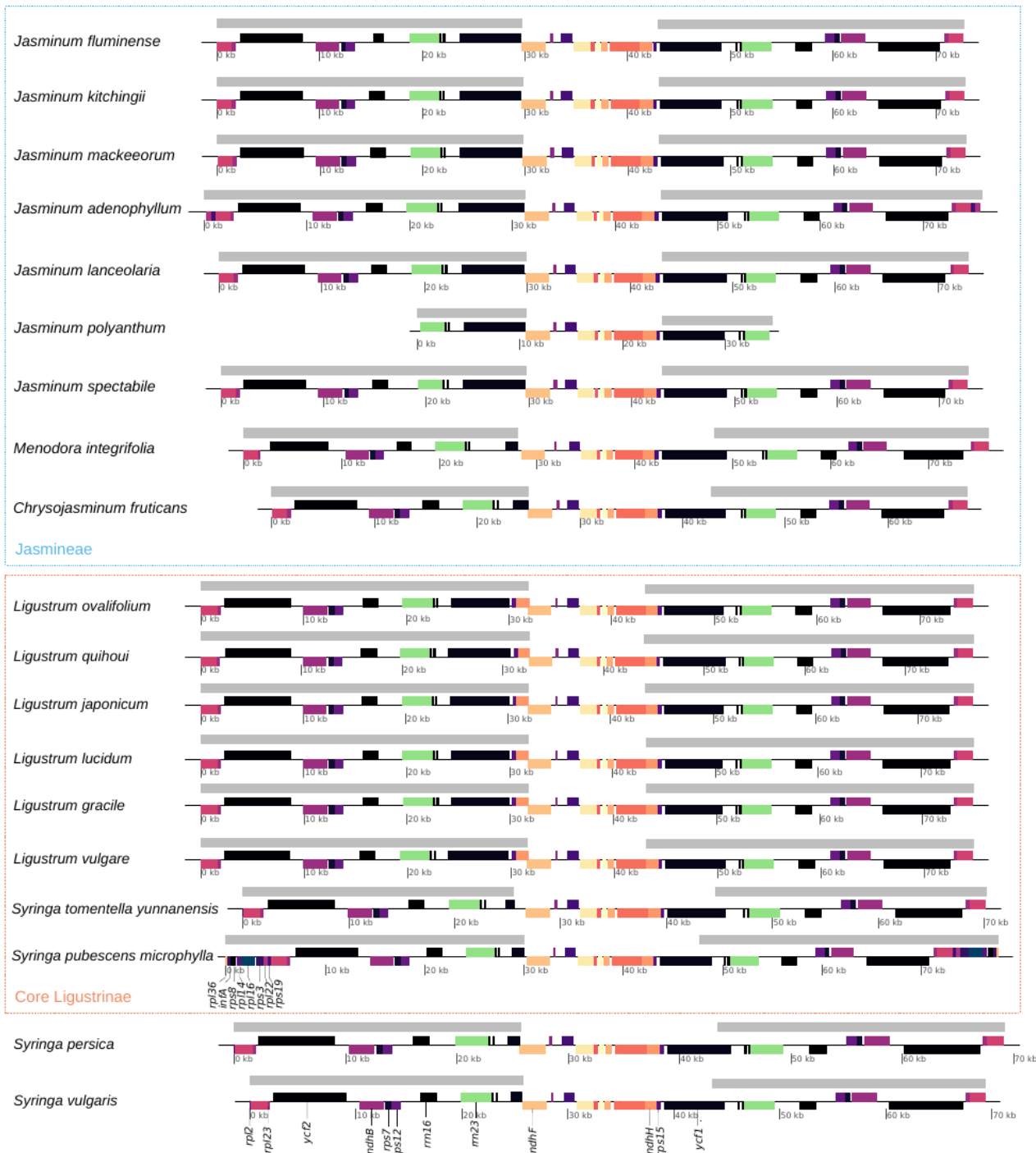


Figure S2. Physical maps summarizing inverted repeat boundary shifts detected in Oleaceae. Each gene is symbolized by a colored block proportional to its size. Grey blocks at the top indicate IR locations. Most Oleaceae exhibit a structure similar to *Syringa persica* and *S. vulgaris*: IR/LSC boundaries are located between *rpl2* and *rps19* and IR/SSC boundaries are located within *ycf1*. Important shifts (implying complete genes) were only observed in Jasmineae and Core Ligustrinae. In Jasmineae, the IRs have extended on the LSC in *J. adenophyllum* to include *rps19* and *rpl22*. In *Jasminum*, IRs have also expanded over the SSC to include the complete sequence of *ycf1*, but not *rps15* and *ndhH*. In *J. polyanthum*, shrinkage of the IRs (of about 15 kb on the IR/LSC boundary) has occurred and they only contain *ycf1* and ribosomal RNA genes. In *Ligustrum*, *ycf1* is completely included in IRs which also contain *rps15* and *ndhH*. In *S. pubescens microphylla*, the IRs have expanded by about 5 kb, encompassing eight protein-coding genes (*rps19*, *rpl22*, *rps3*, *rpl16*, *rpl14*, *rps8*, *infA*, *rpl36*) usually located in the LSC. Abbreviations: IR = Inverted Repeat; SSC = Short Single Copy; LSC = Long Single copy.

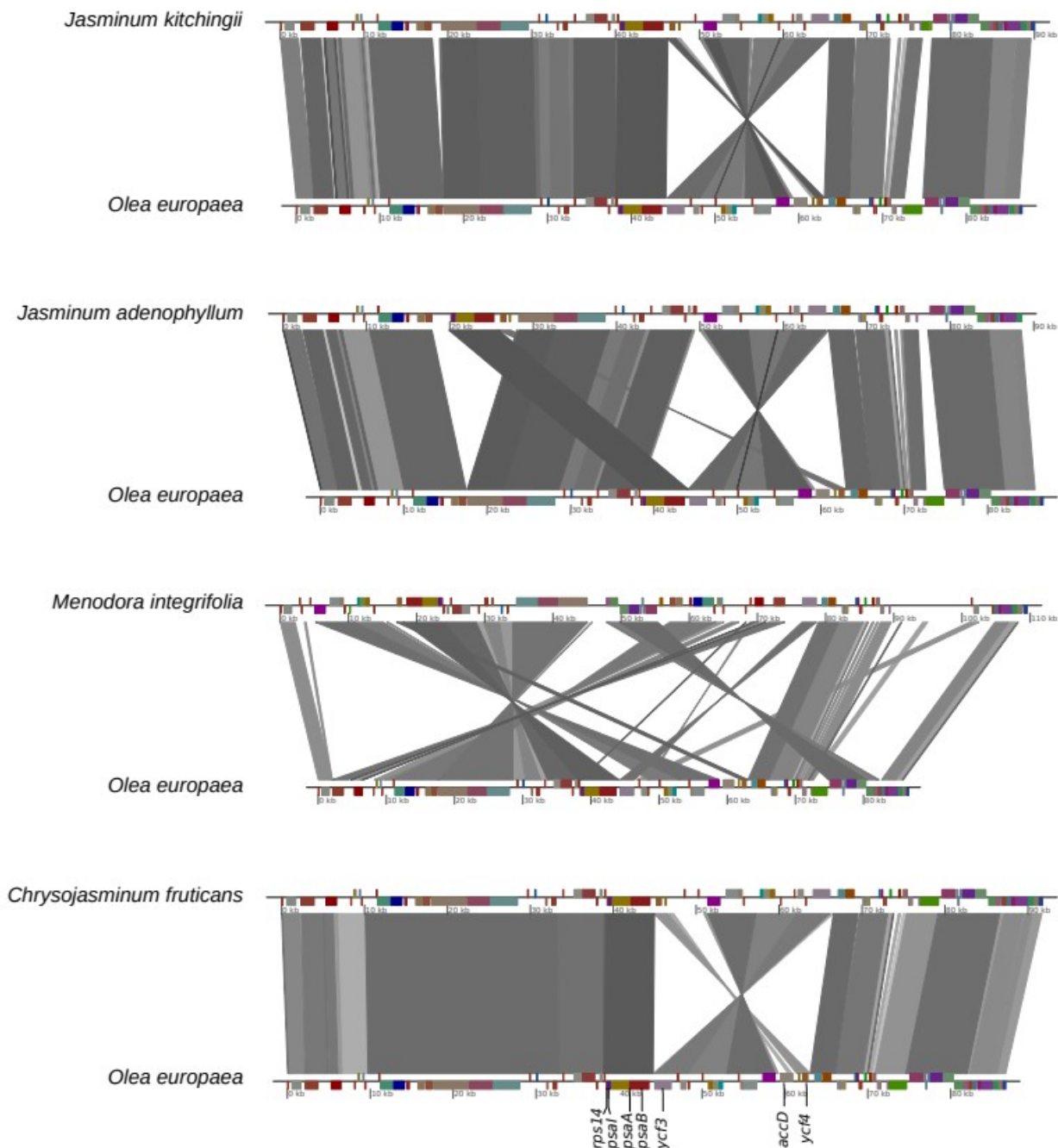
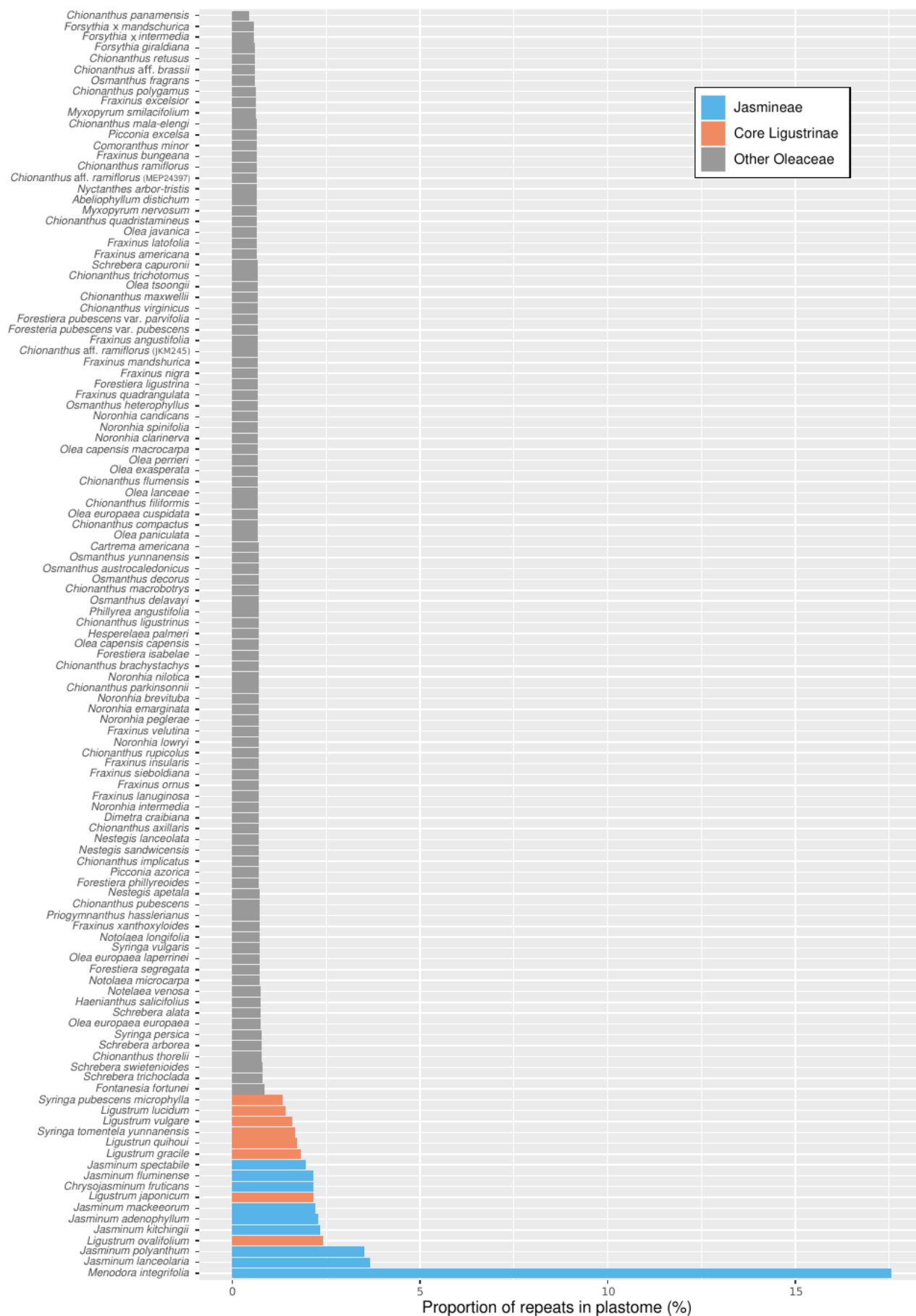


Figure S3. Physical maps summarizing rearrangements detected in the long single copy (LSC) of Jasmineae plastomes. Reorganizations are presented relatively to the plastome of *O. europaea* that shares the same organization than other species belonging to tribes Myxopyreae, Fontanesieae, Forsythieae and Oleaeae (Oleaceae), as well as *Solanum* (Solanaceae), suggesting an ancestral state for these taxa. Each gene is symbolized by a colored block proportional to its size. Three different organizations were detected in Jasmineae. All investigated species of *Chrysojasminum* and *Jasminum* clearly exhibit the same inversion between *ycf4* and *ycf3*, suggesting a single event in their common ancestor. *C. fruticans* and *J. kitchingii* were picked as examples to illustrate this organization. Only *J. adenophyllum* displays a second change in gene order in what appears to be a translocation event of a fragment including *ycf4-psaI-psaB-psaA-rps14*. *Menodora integrifolia* plastome is highly reorganized and the most parsimonious scenario inferred with GRIMM predicts that seven successive inversion events are required to get to this organization starting from the canonical organization in Oleaceae plastome.



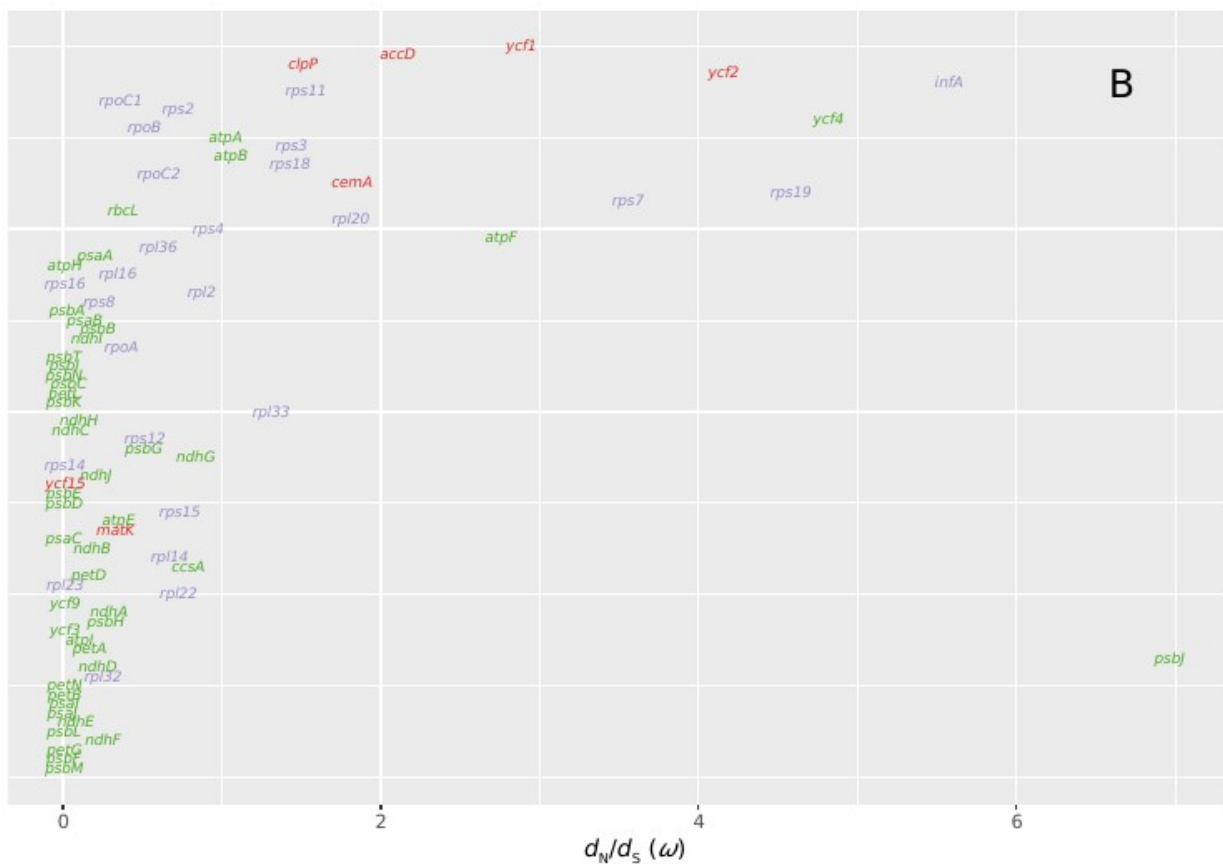
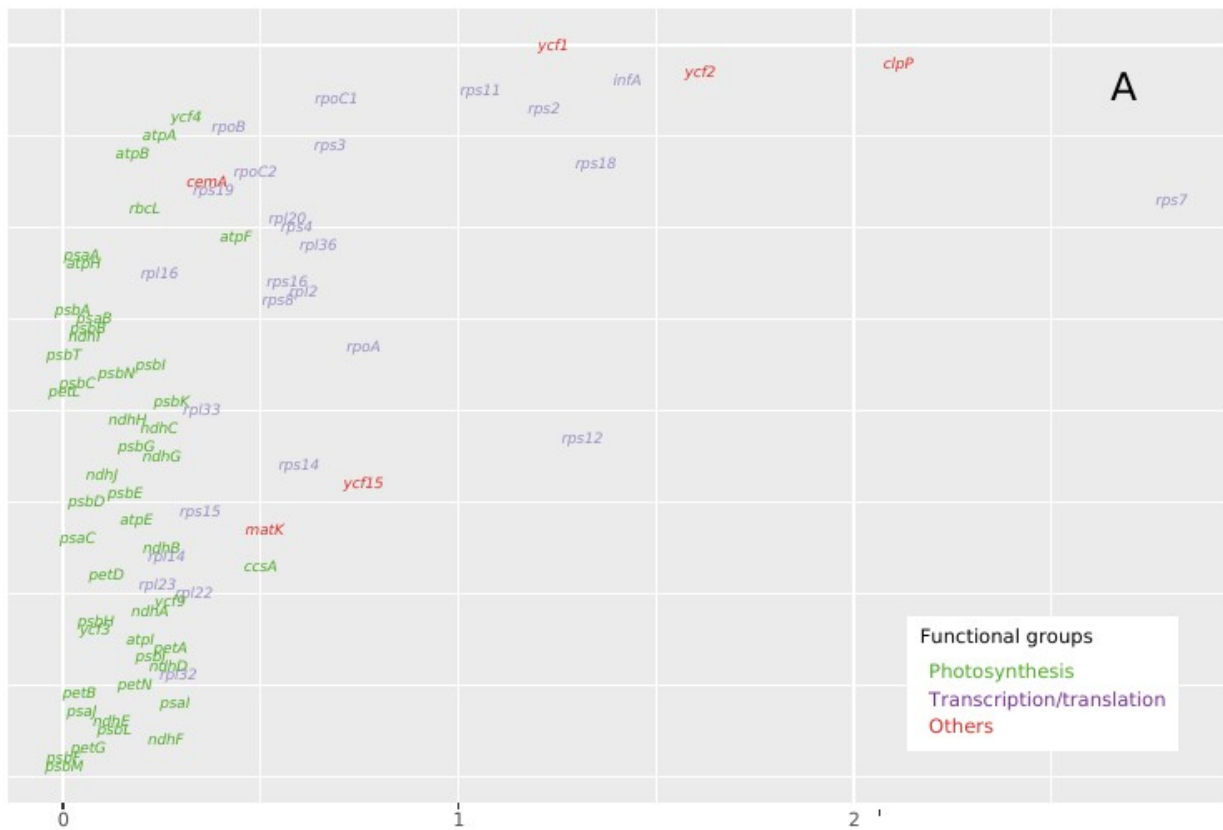


Figure S5. Omega values for plastid coding genes in Jasmineae (A) and Core Ligustrinae (B). Omegas are optimized values under the branch model allowing three distinct ratios for Jasmineae, Core Ligustrinae and other Oleaceae.

Table S1. Gene and intron losses detected in Oleaceae.**Protein-coding genes****Taxa**

	<i>Other Oleaceae</i>	<i>C. fruticans</i>	<i>J. polyanthum</i>	<i>J. fluminense</i>	<i>J. kitchingii</i>	<i>M. integrifolia</i>	<i>J. mackeeorum</i>	<i>J. lanceolaria</i>	<i>J. adenophyllum</i>	<i>L. ovalifolium</i>
<i>accD</i>	+	xb	x	x	x	xb	x	x	x	+
<i>clpP</i> introns	+	x	x	x	x	x	x	x	x	+
<i>psaI</i>	+	+	+	+	+	x	+	+	+	+
<i>rpl23</i>	+	+	+	+	+	xa	+	+	+	+
tRNA genes										
<i>trnA</i> -UGC	2	-	1c	-	-	-	-	-	-	-
<i>trnR</i> -UCU	1	-	-	-	-	-	-	-	-	-
<i>trnR</i> -ACG	2	-	-	-	-	-	-	-	-	-
<i>trnN</i> -GUU	2	-	-	-	-	-	-	-	-	-
<i>trnN</i> -GUC	1	-	-	-	-	-	-	-	-	-
<i>trnC</i> -GCA	1	-	-	-	-	-	-	-	-	-
<i>trnQ</i> -UUG	1	-	-	-	-	-	-	-	-	-
<i>trnE</i> -UUC	1	-	-	-	-	-	-	-	-	-
<i>trnG</i> -GCC	1	-	-	2	-	-	2	2	-	-
<i>trnH</i> -GUG	1	-	-	-	-	-	-	-	-	-
<i>trnI</i> -GAU	2	0	1c	-	-	-	-	-	-	-
<i>trnL</i> -CAA	2	-	1c	-	-	-	-	-	-	-
<i>trnL</i> -UAG	1	-	-	-	-	-	-	-	-	-
<i>trnM</i> -CAU	4	5d	-	-	5d	10d,e	5d	5d	5d	-
<i>trnF</i> -GAA	1	-	-	-	-	-	-	-	-	-
<i>trnP</i> -UGG	1	-	-	-	-	-	-	-	-	-
<i>trnS</i> -GCU	1	2	-	-	-	-	-	-	-	-
<i>trnS</i> -GGA	1	-	-	-	-	-	-	-	-	-
<i>trnS</i> -UGA	1	-	-	-	-	-	-	-	-	-
<i>trnT</i> -GGU	1	-	-	-	-	-	-	-	-	2
<i>trnT</i> -UGU	1	-	-	-	-	-	-	-	-	-
<i>trnW</i> -CCA	1	-	-	-	-	-	-	-	-	-
<i>trnY</i> -GUA	1	-	-	-	-	-	-	-	-	-
<i>trnV</i> -GAC	2	-	1c	-	-	-	-	-	-	-

For protein-coding genes, + denotes presence, x absence; For tRNA genes, - indicates that the number of copy is identical to that of other Oleaceae.

^a In-frame stop codon; ^b 3' part still identifiable (including well-conserved catalytic site); ^c linked with IR boundary shifts; ^d extra copy upstream *accD* fragment, highest identity with IR copy; ^e five other duplicates (one in SSC, two in IRs, and two in LSC) more similar to the copy near *rps14*.

Table S2. cIPp exonic and intronic raw pairwise distance with *Olea europaea*.

Taxa	Exons	Introns	Taxa	Exons	Introns
<i>Abeliophyllum distichum</i>	0.010	0.008	<i>Jasminum lanceolaria</i>	0.466	NA
<i>Cartrema americana</i>	0.005	0.005	<i>Jasminum mackeeorum</i>	0.475	NA
<i>Chionanthus axillaris</i>	0.007	0.007	<i>Jasminum polyanthum</i>	0.469	NA
<i>Chionanthus brachystachys</i>	0.007	0.007	<i>Jasminum spectabile</i>	0.469	NA
<i>Chionanthus aff. brassii</i>	0.007	0.008	<i>Ligustrum gracile</i>	0.231	0.016
<i>Chionanthus compactus</i>	0.005	0.009	<i>Ligustrum japonicum</i>	0.229	0.016
<i>Chionanthus filiformis</i>	0.005	0.010	<i>Ligustrum lucidum</i>	0.243	0.017
<i>Chionanthus fluminensis</i>	0.007	0.010	<i>Ligustrum ovalifolium</i>	0.191	0.022
<i>Chionanthus implicatus</i>	0.005	0.008	<i>Ligustrum quihoui</i>	0.204	0.016
<i>Chionanthus ligustrinus</i>	0.003	0.011	<i>Ligustrum vulgare</i>	0.197	0.014
<i>Chionanthus macrobotrys</i>	0.009	0.012	<i>Myxopyrum nervosum</i>	0.019	0.031
<i>Chionanthus mala-elengi</i>	0.007	0.004	<i>Myxopyrum smilacifolium</i>	0.017	0.032
<i>Chionanthus maxwellii</i>	0.007	0.007	<i>Nestegis apetala</i>	0.009	0.005
<i>Chionanthus panamensis</i>	0.007	0.008	<i>Nestegis lanceolata</i>	0.009	0.004
<i>Chionanthus parkinsonii</i>	0.007	0.004	<i>Nestegis sandwicensis</i>	0.009	0.003
<i>Chionanthus polygamus</i>	0.010	0.008	<i>Noronhia brevitiba</i>	0.010	0.004
<i>Chionanthus pubescens</i>	0.007	0.007	<i>Noronhia candicans</i>	0.010	0.006
<i>Chionanthus quadristamineus</i>	0.007	0.007	<i>Noronhia clarinerva</i>	0.010	0.006
<i>Chionanthus ramiflorus</i>	0.009	0.007	<i>Noronhia emarginata</i>	0.009	0.006
<i>Chionanthus aff. ramiflorus (JKM245)</i>	0.007	0.005	<i>Noronhia intermedia</i>	0.010	0.005
<i>Chionanthus aff. ramiflorus (MEP24397)</i>	0.007	0.005	<i>Noronhia lowryi</i>	0.009	0.005
<i>Chionanthus retusus</i>	0.009	0.005	<i>Noronhia nilotica</i>	0.005	0.008
<i>Chionanthus rupicolus</i>	0.010	0.006	<i>Noronhia peglerae</i>	0.009	0.004
<i>Chionanthus thorelii</i>	0.009	0.004	<i>Noronhia spinifolia</i>	0.010	0.007
<i>Chionanthus trichotomus</i>	0.007	0.011	<i>Notelaea longifolia</i>	0.012	0.007
<i>Chionanthus virginicus</i>	0.007	0.006	<i>Notelaea microcarpa</i>	0.012	0.007
<i>Chrysojasminum fruticans</i>	0.302	0.036	<i>Notelaea venosa</i>	0.012	0.007
<i>Comoranthus minor</i>	0.009	0.005	<i>Nyctanthes arbor-tristis</i>	0.023	0.036
<i>Dimetra craibiana</i>	0.026	0.040	<i>Olea capensis capensis</i>	0.003	0.002
<i>Fontanesia fortunei</i>	0.010	0.015	<i>Olea capensis macrocarpa</i>	0.003	0.003
<i>Forestiera isabelae</i>	0.002	0.009	<i>Olea europaea</i>	0.000	0.000
<i>Forestiera ligustrina</i>	0.002	0.006	<i>Olea europaea cuspidata</i>	0.000	0.001
<i>Forestiera phillyreoides</i>	0.002	0.007	<i>Olea europaea laperrinei</i>	0.000	0.001
<i>Forestiera pubescens var. parviflora</i>	0.002	0.007	<i>Olea exasperata</i>	0.003	0.002
<i>Forestiera pubescens var. pubescens</i>	0.002	0.007	<i>Olea javanica</i>	0.009	0.007
<i>Forestiera segregata</i>	0.002	0.007	<i>Olea lancea</i>	0.003	0.002
<i>Forsythia giraldiana</i>	0.010	0.011	<i>Olea paniculata</i>	0.005	0.003
<i>Forsythia x intermedia</i>	0.010	0.011	<i>Olea perrieri</i>	0.003	0.002
<i>Forsythia x mandshurica</i>	0.010	0.011	<i>Olea tsoongii</i>	0.009	0.008
<i>Fraxinus americana</i>	0.009	0.011	<i>Osmanthus austrocaledonicus</i>	0.009	0.003
<i>Fraxinus angustifolia</i>	0.005	0.008	<i>Osmanthus decorus</i>	0.009	0.004
<i>Fraxinus bungeana</i>	0.007	0.008	<i>Osmanthus delavayi</i>	0.009	0.006
<i>Fraxinus excelsior</i>	0.005	0.008	<i>Osmanthus fragrans</i>	0.009	0.006
<i>Fraxinus insularis</i>	0.005	0.008	<i>Osmanthus heterophyllus</i>	0.009	0.005
<i>Fraxinus lanuginosa</i>	0.005	0.009	<i>Osmanthus yunnanensis</i>	0.009	0.005
<i>Fraxinus latifolia</i>	0.009	0.010	<i>Phillyrea angustifolia</i>	0.009	0.004
<i>Fraxinus mandshurica</i>	0.007	0.007	<i>Picconia azorica</i>	0.009	0.004
<i>Fraxinus nigra</i>	0.005	0.007	<i>Picconia excelsa</i>	0.010	0.005
<i>Fraxinus ornus</i>	0.005	0.009	<i>Priogymnanthus hasslerianus</i>	0.010	0.011
<i>Fraxinus quadrangulata</i>	0.005	0.008	<i>Schrebera alata</i>	0.009	0.006
<i>Fraxinus sieboldiana</i>	0.005	0.009	<i>Schrebera arborea</i>	0.009	0.007
<i>Fraxinus velutina</i>	0.009	0.010	<i>Schrebera capuronii</i>	0.009	0.007
<i>Fraxinus xanthoxylodes</i>	0.007	0.007	<i>Schrebera swietenoides</i>	0.010	0.008
<i>Haenianthus salicifolius</i>	0.003	0.006	<i>Schrebera trichoclada</i>	0.009	0.006
<i>Hesperelaea palmeri</i>	0.002	0.004	<i>Syringa persica</i>	0.009	0.014
<i>Jasminum adenophyllum</i>	0.472	NA	<i>Syringa pubescens microphylla</i>	0.152	0.019
<i>Jasminum fluminense</i>	0.472	NA	<i>Syringa tomentella yunnanensis</i>	0.064	0.013
<i>Jasminum kitchingii</i>	0.472	NA	<i>Syringa vulgaris</i>	0.007	0.012

Table S3. List of Oleaceae species for which the inheritance mode of plastids has been reported.

Tribe, Subtribe	Taxon	Plastid inheritance ^a	
		Mode of inheritance	References ^c
Oleaceae, Oleinae	<i>Olea europaea</i> ^b	Maternal	[5]
	<i>Chionanthus retusus</i>	Maternal	[4]
Oleaceae, Fraxininae	<i>Fraxinus excelsior</i>	Maternal	[4]
Oleaceae, Ligustrinae			
Genus Syringa, Subgenus Syringa, Series Syringa	<i>Syringa oblata</i>	Maternal	[2,4]
	<i>Syringa vulgaris</i>	Maternal	[1,2]
	<i>Syringa chinensis</i>	Maternal	[2]
	<i>Syringa persica</i>	Maternal	[2]
	<i>Syringa protolaciniata</i>	Maternal	[2]
	<i>Syringa dilata</i>	Maternal	[2]
Genus Syringa, Subgenus Syringa, Series Pinnatifoliae	<i>Syringa pinnatifolia</i>	Maternal	[2]
Genus Syringa, Subgenus Syringa, Series Villosae	<i>Syringa wolfii</i>	Paternal leakage	[2]
	<i>Syringa sweginzowii</i>	Paternal leakage	[2]
	<i>Syringa yunnanensis</i>	Paternal leakage	[2]
	<i>Syringa villosa</i>	Paternal leakage	[2]
	<i>Syringa josikaea</i>	Paternal leakage	[2]
Genus Syringa, Subgenus Syringa, Series Pubescens	<i>Syringa meyeri</i>	Paternal leakage	[2]
	<i>Syringa pubescens microphylla</i>	Paternal leakage	[2]
	<i>Syringa patula</i>	Paternal leakage	[2]
Genus Syringa, Subgenus Ligustrina	<i>Syringa pekinensis</i>	Paternal leakage	[2,4]
	<i>Syringa reticulata var amurensis</i>	Paternal leakage	[2]
	<i>Syringa reticulata</i>	Paternal leakage	[2]
Genus Ligustrum	<i>Ligustrum lucidum</i>	Paternal leakage	[2]
	<i>Ligustrum quihoui</i>	Paternal leakage	[2]
	<i>Ligustrum vulgare</i>	Paternal leakage	[2]
Jasmineae	<i>Jasminum officinale</i>	Paternal leakage	[3]
	<i>Jasminum nudiflorum</i>	Paternal leakage	[3]
	<i>Jasminum polyanthum</i>	Paternal leakage	[4]
Fontanesieae	<i>Fontanesia fortunei</i>	Maternal	[4]

^aThe inheritance mode (i.e. maternal or potential paternal leakage) is determined by epifluorescence microscopic observation of organellar DNA in mature pollen; ^bTaxa in bold were included in the present work; ^cReferences: [1] Corriveau and Coleman (1988); [2] Liu et al. (2004); [3] Sodmergen et al. (1998); [4] Zhang et al. (2003); [5] Hagemann and Schröder (1989).

References

- Corriveau JL, Coleman AW. 1988. Rapid screening method to detect potential biparental inheritance of plastid DNA and results for over 200 angiosperm species. *Am J Bot.* 75:1443–1458.
- Liu Y, Cui H, Zhang Q, Sodmergen. 2004. Divergent potentials for cytoplasmic inheritance within the genus *Syringa*. A new trait associated with speciation. *Plant Physiol.* 136:2762–2770.
- Sodmergen, Bai HH, He JX, Kuroiwa H, Kawano S, Kuroiwa T. 1998. Potential for biparental cytoplasmic inheritance in *Jasminum officinale* and *Jasminum nudiflorum*. *Sex Plant Reprod.* 11:107–112.
- Zhang Q, Liu Y, Sodmergen. 2003. Examination of the cytoplasmic DNA in male reproductive cells to determine the potential for cytoplasmic inheritance in 295 angiosperm species. *Plant Cell Physiol.* 44:941–951.

Chapter 3

A hemizygous supergene controls homomorphic di-allelic self-incompatibility in olive

Raimondeau P., Marande W., Cheptou P.-O., Vautrin S., Manzi S.,
Chave J., Christin P.-A., Besnard G.

Abstract

Most angiosperms are hermaphrodites and 75% even exhibit male and female organs on the same floral structure. This organization leads to an increased risk of self-fertilization that may generate inbreeding depression. Mechanisms to prevent selfing evolved at least 35 times independently across plant families. In *Olea europaea* (olive), a unique system with only two incompatibility groups of morphologically identical individuals has been described but the underlying genetic determinism has not been elucidated. By comparing genomic profiles of wild and cultivated olive individuals belonging to the two incompatibility groups, in three subspecies, we show that the *S*-locus in olive is a 700-kb hemizygous region present in only one of the two mating types. This genomic architecture is similar to that of the supergenes controlling distyly in other lineages. This co-occurrence is consistent with the di-allelic nature of the two systems and reinforces the interest of the Oleaceae as a model for studying the evolution of self-incompatibility systems.

Key-words: Haplotype-resolved genome assembly, Heterostyly, *Olea europaea*, Oleaceae, *S*-locus.

Introduction

Self-incompatibility (SI) is the most common mechanism promoting outbreeding in flowering plants and represents a remarkable example of convergent evolution having evolved independently at least 35 times (Igic *et al.*, 2008). In self-incompatible plants, pollen–pistil recognition systems prevent self-fertilization but also cross-fertilization between individuals sharing the same compatibility type. Depending on its association with morphological variation between incompatibility groups, self-incompatibility can be deemed heteromorphic or homomorphic. At the genetic level, there are two major classes of systems: if the incompatibility phenotype of the pollen is determined by the gametophyte haploid genome, the system is called gametophytic, and if it is determined by the genome of the parent sporophyte, the system is called sporophytic. To date, few self-incompatibility systems have been characterized at the molecular level (see Fujii *et al.*, 2016 for a review). The molecular effectors vary but are generally encoded by a single genomic region, the *S*-locus, containing several genes, encoding pollen and pistil components of the self-incompatibility reaction.

In olive (*Olea europaea* L., Oleaceae), self-incompatibility is a major obstacle for olive productivity as it restricts varietal choice, olive breeding and can hinder pollination effectiveness. Moreover, as no morphological differences enables the identification of incompatibility groups, orchard are composed empirically. The elucidation of olive self-incompatibility mechanism is consequently of great interest for olive production. Major advances have been made in the past few

years towards this goal. First, examinations of cross-compatibilities revealed the existence of a peculiar sporophytic self-incompatibility system with only two incompatibility groups (Saumitou-Laprade *et al.*, 2017; Besnard *et al.*, 2020). It has been hypothesized that two alleles, *S* and *s*, exist at the postulated diallelic self-incompatibility (DSI) locus, with *S* dominant over *s*, and with only two genotypic combinations (*Ss* and *ss*), corresponding to the G1 and G2 incompatibility groups, respectively (Saumitou-Laprade *et al.*, 2010). This configuration is unparalleled in plants, as self-incompatibility systems generally gather numerous alleles. For instance, between 30 and 40 *S*-alleles are typically found within natural populations of Brassicaceae and Asteraceae (Lawrence, 2000). Genetic mapping of the DSI locus in a biparental population revealed a region of 600 kb linked to the self-incompatibility phenotype on chromosome 18 of the reference oleaster genome assembly published by Unver *et al.* (2017), but also suggested that the region was probably misassembled (Mariotti *et al.*, 2020). This region encompasses 17 genes, but no obvious candidate for a role in self-incompatibility was identified.

In addition to its importance for olive production, the uniqueness of this homomorphic DSI system strengthens the interest in deciphering it. Indeed, negative frequency-dependent selection is expected to promote the emergence of new *S*-alleles (Saumitou-Laprade *et al.*, 2010). The maintenance of only two alleles is even more intriguing when looked at the family scale. In Oleaceae, homomorphic DSI have been identified in other distantly related genera in the Oleae tribe; *Phillyrea* (Saumitou-Laprade *et al.*, 2010), *Fraxinus* (Vernet *et al.*, 2016) and *Ligustrum* (de Cauwer *et al.*, 2020). Cross-species pollination experiments even demonstrated that recognition specificities are shared between these divergent lineages (Vernet *et al.*, 2016; Saumitou-Laprade *et al.*, 2017). It therefore suggests that this specific self-incompatibility system has been conserved for a period of time of at least 40 Ma (Olofsson *et al.*, 2019), probably thanks to a genetic architecture preventing recombination. Genetic mapping of the DSI locus in *Phillyrea angustifolia* identified several associated markers, including one in a region homologous to the candidate region on olive chromosome 18, even if it also pointed out assembly errors of olive chromosome 18 (Carré *et al.*, 2021). Finally, another open question about this homomorphic DSI concerns its origin. The heteromorphic DSI system is considered ancestral in Oleaceae and is encountered in other tribes of the family (Taylor, 1945). The genetic controls are unknown but it has been hypothesized that the two DSI systems might be homologous (Saumitou-Laprade *et al.*, 2010). Solving these puzzling questions over the origin, evolution and maintenance of the homomorphic DSI requires first elucidating its genetic determinism.

In this study, we took advantage of three reference genome assemblies for cultivated and wild olives, including a newly generated haplotype-resolved assembly of the Saharan wild olive

[*Olea europaea* subsp. *laperrinei* (Batt. & Trab.) Cif.]. We used RAD-sequencing of individuals from the two incompatibility groups, from three olive subspecies, to detect genomic regions associated with self-incompatibility. We propose a new hypothesis on the genomic location of the *S*-locus in olive as we demonstrate that a hemizygous region is present in only one of the two mating types, suggesting it controls self-incompatibility. Our results are the first uncovering the implication of a hemizygous region in homomorphic self-incompatibility. This finding is consistent with the diallelic nature of self-incompatibility in olive and mirrors the uniqueness of the homomorphic system in the family.

Material and Methods

WGS and assembly of the Laperrine's olive genome

High-molecular weight DNA from fresh leaves was extracted for whole genome sequencing of Laperrine's olive individual 'Adjelella 9_S4' (phenotyped as G1, and thus expected to have a *Ss* genotype; Besnard *et al.*, 2020). Sequencing was performed using two PacBio HiFi SMRT cells. HiFi reads were corrected and assembled using HiFiasm (Cheng *et al.*, 2021) which produces partially phased assemblies, i.e. two complete assemblies with long stretches of phased blocks, representing an entire diploid genome. Assembly completeness was assessed with BUSCO against eudicots_odb10 (Simão *et al.*, 2015) and allele separation was checked via *k*-mer analysis with KAT (Mapleson *et al.*, 2017). All steps from extraction to assembly were performed at the CNRGV (Centre National de Ressources Génomiques Végétales, Toulouse, France). We used Minimap2 (Li, 2018) to map back long read on the haplotype assembly and to perform whole-chromosome alignment. Genome assemblies were softmasked using Red (Girgis, 2015) and annotated using Braker2 (Brůna *et al.*, 2021) with hints from OrthoDB (Zdobnov *et al.*, 2021). We only kept genes at least partially supported by external hints. Functional annotation was performed with InterProScan with default parameters (Jones *et al.*, 2014). We used EDTA (Ou *et al.*, 2019) to identify transposable elements (TE). Syntenic regions between the three olive genome assemblies were identified using MCscan and the jvarkit python library (Tang *et al.*, 2008).

Plant sampling, RAD-sequencing and reads processing

We selected 37 accessions of three different olive subspecies (9 *O. e.* subsp. *europaea*, 18 *O. e.* subsp. *laperrinei*, and 10 *O. e.* subsp. *cuspidata*) for which the SI phenotype has been previously determined using paternity tests on realized matings (Besnard *et al.*, 2020). In total, 19 individuals are G1 (*Ss*) and 18 are G2 (*ss*; Table S1). DNAs were extracted with BioSprint (Qiagen), and 200 ng were then digested with *Pst*I. RAD-seq libraries were prepared at the GenoToul sequencing

platform facility (Toulouse, France) following the protocol described in Etter *et al.* (2011). Samples were pooled into three distinct libraries and multiplexed. The three libraries were sequenced on one lane of Illumina NovaSeq run to produce 150-bp paired-end reads. Reads were demultiplexed and cleaned with the `process_radtags` module of Stacks v2.5 (Rochette *et al.*, 2019), allowing the rescue of reads with two mismatches in barcodes (less than the distance between any two barcodes in the used set). Cleaned demultiplexed reads were then mapped using Bowtie2 (with default parameters; Langmead and Salzberg, 2012) to the oleaster reference genome (Unver *et al.*, 2017), as well as the reference genome for the olive cultivar ‘Arbequina’ (Rao *et al.*, 2021), and finally to the haplotype genome assemblies of the Laperrine’s olive generated in this study.

Variant calling and population structure

From read alignments on each reference genome, we called bi-allelic SNPs with `bcftools` (Danecek *et al.*, 2021), keeping only reads with a mapping quality above 20 (excluding multi-mappers), a base-calling quality of at least 20, and a mean read depth over all individuals between 5 and 60 (twice the mean depth to exclude errors due to mapping in paralogous/repetitive regions). We further filtered sites missing in more than 90% of individuals and those with a minor allele frequency inferior to 0.05 using `vcftools` (Danecek *et al.*, 2021). For an individual genotype to be called, we also required a minimum coverage of 5. We only kept one site every 1000 bp to perform a principal component analysis (PCA) of the 37 samples using `vcfR` and `adegenet` packages in R (Jombart and Ahmed, 2011; Knaus and Grünwald, 2017).

Genetic differentiation between incompatibility groups

Given that incompatibility groups are functionally conserved in divergent Oleaceae lineages (Vernet *et al.*, 2016), one could expect the *S*-locus to be highly divergent between the two no-recombining haplotypes (Kamau and Charlesworth, 2005). Pairwise F_{ST} were calculated in 50-kb sliding-windows along the different reference genomes with `vcftools`. To test whether the differentiation between incompatibility group deviates from null expectations, we performed 1,000 permutations by randomly shuffling individuals among the two groups before re-estimating F_{ST} for each window. We then calculated *p*-values as the proportion of permuted F_{ST} values that were larger than the observed one, applying a Benjamini-Hochberg correction for multiple testing with the `p.adjust` function in R.

Detecting variation in depth of coverage between groups

The two haplotypes may have diverged enough to prevent reads from one haplotype to properly map to the haplotype assembled in the haploid reference, similarly to what is observed for sex chromosomes (Qiu *et al.*, 2016; Palmer *et al.*, 2019). Moreover, hemizyosity is a common feature of several heteromorphic DSI systems [e.g. *Primula* (Li *et al.*, 2016), *Turnera* (Shore *et al.*, 2019), *Linum* (Gutiérrez-Valencia *et al.*, 2022)] and could explain the absence of recombination permitting the conservation of the DSI system over a long evolutionary period. Both configurations would result in sequencing-depth variation between groups. The cross-compatibility phenotype of the individual sequenced for the oleaster reference genome is unknown. Both cultivar ‘Arbequina’ and the Laperrine’s individual we used for whole-genome sequencing are G1 (Besnard *et al.*, 2020), and thus expected to be *Ss*. If the assembled haplotype is *s*, the *Ss:ss* ratio would be close to 0.5 at the *S*-locus. Conversely, if the assembled haplotype is *S*, the *Ss:ss* ratio would be infinite. We used SeqKit (Shen *et al.*, 2016) to predict *PstI* restriction sites in the three different reference genomes to which reads were mapped. We then estimated the depth of coverage at each predicted *PstI* locus with samtools (Danecek *et al.*, 2021). For each sample, we normalized the counts by dividing by the total number of mapped reads. We filtered out sites, removing those with a median normalized coverage smaller than 0.5 read per million of mapped reads across both groups. The ratio between the median depth in *Ss* individuals and *ss* individuals, calculated as $\log_2[(Ss+1)/(ss+1)]$, along the genome was plotted by 100-kb sliding-windows with a step size of 25 kb, using the python library matplotlib (Hunter, 2007). We also examined the occurrence of a significant depth difference between *Ss* and *ss* individuals in 50 kb windows along each reference genome using two-sided Wilcoxon tests. To increase our power and as we are looking for a genetic region conserved across species, we pooled individuals from each incompatibility group together, regardless of the olive subspecies.

Results

High-quality haplotype assemblies of the Saharan olive genome

The two SMRT cells produced 3,659,295 reads with a median size of 18.75 kb. The *k*-mer model estimated a coverage of 48× for a haploid genome size of 1.3 Gb. Hifiasm produced two partially phased assemblies. These are complete and aim at representing each haplotype of a diploid genome. However, some switching can happen between the two haplotypes and chromosomes are not phased (Cheng *et al.*, 2021). The assembly for ‘haplotype 1’ consists of 239 contigs with a N50 of 52.5 Mb for a total length of 1.4 Gb. It has 1.2% fragmented and 97.4% complete BUSCO eudicots genes, of which 18.8% are duplicated. For ‘haplotype 2’, the assembly consists of 154 contigs with a N50 of 32.9 Mb for a total length of 1.3 Gb. In this second assembly, 96.3% of BUSCO eudicots genes are complete, 18.2% are present as duplicates, and 1.3% are fragmented. The proportion of duplicated

genes is congruent with what is expected and observed in other Oleaceae species due to the allopolyploid origin of the tribe (Zhang *et al.*, 2020). A *k*-mer analysis confirms good allele separation between the two assemblies (Figure S1). Overall, 68,607 and 65,939 protein-coding genes were annotated in haplotypes 1 and 2 respectively, with at least partial support from protein hints. Genomes are composed of about 44% of repeats, the most abundant types being LTR transposable elements (80% of repeats) and DNA transposons (15%), consistent with reports for the wild Mediterranean olive genome (Unver *et al.*, 2017). Synteny analyses identified 24 scaffolds collinear with the 23 chromosomes of the previously published assembly in haplotype 1, as chromosome 18 is split into two scaffolds in our assembly (Figure 1). For haplotype 2, we retrieved 28 scaffolds that match the original 23 chromosomes. Our assemblies are largely collinear with the ‘Arbequina’ reference genome, indicating the genomic structure has been stable within the species at odds with a previous report (Julca *et al.*, 2018).

Olive subspecies are strongly-structured based on nuclear SNP data

Across the 37 olive samples, a total of 618,291,848 paired end 150-bp reads were generated, 99% of which were retained after quality filtering and demultiplexing (Table S2). The mapping rate over the wild olive genome assembly ranged from 74 to 85% per sample and was higher for the ‘Arbequina’ genome (78.5 to 90%) and for both haploid assemblies of the Laperrine’s olive (84 to 93.5%). We used a set of about 40,000 bi-allelic SNPs to study the genetic structure of the three subspecies. The PCA clearly clustered the samples into three distinct groups corresponding to each of the studied subspecies (Figure S2). There was a greater diversity in our Mediterranean olive sampling set than in the two other subspecies.

No highly-differentiated genomic regions identified between incompatibility groups

We performed F_{ST} scans with our RAD-seq data mapped along each genome assembly to identify windows highly differentiated between the two incompatibility groups (Figure S3). In each genome for each subspecies, genomic scans identified about 500 windows among a dozen thousands investigated for which the observed F_{ST} was larger than expected if individuals were randomly assigned to a group (p -values < 0.05), but none were significant after correction for multiple testing. Being limited by the power of our approach (small number of individuals and large window size due to the fragmented nature of RAD-seq coverage), we still looked at the windows in which the observed F_{ST} was greater than the one observed over the 1000 permutations. We only identified such windows using Laperrine’s olive individuals. Regardless of the reference genome used, four to 11 windows had observed F_{ST} greater than in any permutation. None of them were in the previously

postulated region of chromosome 18, but we did identify F_{ST} peaks at different coordinates on this chromosome in each assembly: one window at 16.45 Mb in the oleaster genome (for a total of 11 regions with high F_{ST}), five windows (at 17.5, 18.6, 19.5, 19.6, and 22.6 Mb) out of nine in the ‘Arbequina’ genome, and three windows out of 11 in ‘haplotype 1’ of the Laperrine’ olive (at 1.6 Mb, and from 2.3 to 2.4 Mb). None of the four windows identified in ‘haplotype 2’ were on chromosome 18.

RAD-seq markers reveal a large peak of differential coverage between incompatibility groups

The four reference genomes (including the two haploid genomes of the Laperrine’s olive) were scanned for variation in coverage ratio between the two incompatibility groups for each subspecies. In the wild olive reference genome, a modest increase in $Ss:ss$ coverage ratio was observed in the region previously associated with the self-incompatibility phenotype (8.5-9.1 Mb on chromosome 18, highlighted in grey on Figure 2; Mariotti *et al.*, 2020), though only using Laperrine’s olive individuals. On the same chromosome, a greater peak, common to the three subspecies was detected between 16.7 and 16.9 Mb. Our Wilcoxon tests for difference in depth between groups (individuals pooled by incompatibility groups regardless of the species), did identify three windows at these coordinates as significantly more covered in Ss individuals ($FDR < 0.05$). Three more windows with this profile were detected on two unanchored scaffolds: NW_019268110.1 (from 0.5 to 1 kb over a total size of 262 kb) and NW_019238463 (0.5 to 1 kb for a total size of 145 kb). We did not detect any windows shared across the three subspecies with the opposite profile (higher coverage in ss than in Ss). Similar results were observed using the olive cultivar ‘Arbequina’ as reference. A unique and large increase in coverage ratio was observed on chromosome 18 in the three subspecies between 18.7 and 19.5 Mb (Figure 2). This was supported by Wilcoxon tests results, as we only detected nine 50-kb windows with significant difference in depth, all between 18.85 and 19.35 Mb. In Laperrine’s olive haplotype assemblies, we detected a large peak in ‘haplotype 1’ between 1.6 and 2.3 Mb (Figure 2). All 50-kb windows between 1.6 and 2.3 Mb were significant for a different depth and were the only significant windows in this assembly. As a consequence, we suggest the assembled haplotype is the dominant S haplotype in both reference genomes as well as in our ‘haplotype 1’ assembly.

In contrast, we did not detect any peak (Figure 2) or significant window ($FDR < 0.05$) in any direction in ‘haplotype 2’, that we infer to be the recessive s haplotype. It means that individuals from both incompatibility types were sequenced and mapped with similar rates on this haplotype assembly. This last observation is consistent with a moderate divergence between the two

haplotypes at the *S*-locus or with the absence of this *S*-locus. As our other observations reject the former, we postulate the dominant *S* haplotype is hemizygous.

A single sequence insertion differentiates the two incompatibility types

Comparison of the scaffolds corresponding to chromosome 18 in the two Laperrine's olive haplotype assemblies (Figure 3) revealed a large indel of over 700-kb underlying the coverage variation between the two groups. This region in 'haplotype 1' has no homolog in 'haplotype2', yet the flanking regions have homologs in 'haplotype2', where they are adjacent (Figure 1). These patterns indicate that a 0.7 Mb-stretch of DNA is present in 'haplotype1', yet missing from 'haplotype2'. We confirmed the existence of this insertion by mapping back HiFi reads on the two haplotype-resolved assemblies. Inspection of the mapped reads at this location confirmed the presence of whole reads covering the junction without the indel (on 'haplotype 2') and other reads including the indel sequence and some downstream/upstream sequence (on 'haplotype 1').

Synteny analysis confirmed that the sequence underlying the large peak observed around 19 Mb in the 'Arbequina' genome is syntenic to the 'haplotype 1'-specific sequence (Figure 1). According to RAD-seq data, this region is totally absent in *ss* individuals, regardless of the subspecies, supporting our conclusion that this sequence is the *S* haplotype. Compared to 'haplotype 1' Laperrine's olive sequence, a deletion of 50 kb within this region is observed in 'Arbequina' (confirmed by RAD-seq data from the same cultivar), which also presents an inversion in the reference genome. In the oleaster assembly, the sequences behind the peak detected on chromosome 18 (at 16.8 Mb) as well as on the two unanchored scaffolds are also syntenic fragments of the 700-kb *S*-specific region. The full *S*-locus is thus present but scattered across this genome assembly. The fragment on chromosome 18 corresponds to the 5' end of the indel sequence, NW_019238463.1 to the middle and NW_019268110.1 to the 3' end, with some downstream sequence. NW_019238463.1 overlaps with one of the scaffolds in *Phillyrea* that showed partial association with DSI in the genetic mapping work of Carré *et al.* (2021). The newly identified region is, in both 'haplotype 1' and 'Arbequina' assemblies, contiguous to the candidate region defined by Mariotti *et al.*, (2020), but 8 Mb apart in the oleaster assembly.

We annotated 17 genes in the 700-kb *S*-specific region of the Laperrine's olive, most of them of unknown functions (Table S3). Five LTRs were identified by EDTA within the region, as well as two MITEs (Figure 3). In addition, four of our predicted protein-coding genes had TE-related functions. Finally, four annotated genes were common in the Laperrine's olive, oleaster, and 'Arbequina' genome annotations. Two of them seem particularly interesting due to their annotation suggesting functions related to hormone regulation: g33223.t1 is annotated as a BES1/BZR1

(BRASSINAZOLE-RESISTANT) transcription factor homolog, and g33233.t1 as a gibberellin 2-beta-dioxygenase homolog. We also note the presence of a second BES/BZR homolog upstream of this one in our annotation, but solely supported by computational prediction and no biological hints. This gene was present in the two other genome annotations.

Discussion

The dominant S haplotype is a hemizygous supergene

Using a combination of already available and newly generated genomics data, we were able to considerably improve the genomic mapping of the olive *S*-locus. We found that the previous candidate region, identified using linkage in a segregating population of cultivated olive (Mariotti *et al.*, 2020), was erroneously associated with self-incompatibility. This was probably caused by assembly errors. The region is contiguous to our newly identified region in all assemblies but the reference genome used by Mariotti *et al.* (Figures 1, 2). The use of wild individuals confers a greater resolution power to our analysis and combining data from distinct subspecies increased our confidence in the region we identified. Our results indicate that a hemizygous region determines DSI in olive. The dominant *S* haplotype harbors a 700-kb region, specific to *Ss* individuals. Hemizyosity of the *S*-locus is supported by coverage ratio analyses between the two incompatibility groups (Figure 2) and by our haplotype-resolved assemblies for the Laperrine's olive (Figure 3).

Theory predicts that the absence of recombination due to a heterozygous state over a long evolutionary time will increase sequence divergence of *S*-locus genes and the accumulation of repetitive elements and transposons (Uyenoyama, 2005). The *S* haplotype is indeed rich in transposable elements and while the statistical power of our analysis was limited, we did detect elevated F_{ST} in the regions surrounding the *S*-locus (Figure S3). This would be consistent with suppressed recombination at a larger scale around the *S*-specific region. Although the centromeres have not been located in olive, the decrease in gene density around 20 Mb on chromosome 18 (Figure S4) could suggest the *S*-locus is pericentromeric, a situation that has been suggested to favor the emergence of a supergene (Kappel *et al.*, 2017). The proximity to centromere and the abundance of repetitive elements could explain the difficulty encountered in the proper assembly of this region, especially with the short-read technologies that were used to generate the oleaster genome assembly (Unver *et al.*, 2017).

The *S*-locus contains multiple genes of unknown functions as well as two genes with predicted functions in regulation of brassinosteroid (BES/BZR) and gibberellin (G2OX). Hormone-mediated signaling plays a central role in plant development. Interestingly, two genes mediating brassinosteroids inactivation were shown to determine female incompatibility in heterostylous *Primula* and *Turnera* (Matzke *et al.*, 2020; Huu *et al.*, 2022). In these two cases, short-styled plants harbor a brassinosteroid inhibitor gene while long-styled plants do not. The *S*-morph pollen consequently depends on the presence of brassinosteroids in the style for tube growth, brassinosteroids that are only produced in L-morph style. On the other hand, the presence of brassinosteroids (due to the absence of inhibitor gene) in L-morph style seems to prevent the elongation of pollen tube of L-morph pollen, but the mechanism is not completely understood. Incompatibility reactions are not perfectly symmetrical in olive, with *Ss* incompatibility reaction happening at stigma surface of *ss* plants, while the reaction with *ss* pollen consists of stopping tube elongation in *ss* style, after germination (Saumitou-Laprade *et al.*, 2017). A self-incompatibility system analogous to the one reported in heterostylous plants could be hypothesized in the case of the olive tree, with *Ss* individuals (bearing the brassinosteroid inhibitor) producing pollen that needs the exogenous brassinosteroids from *ss* style for germination. The absence of another effector might cause the arrested growth of *ss* pollen tube in *ss* style. Further functional studies are required to uncover the precise mechanism encoded at the *S*-locus, and notably, to conclude about the potential role of brassinosteroid regulation in the olive self-incompatibility system.

Conclusions

Our work reveals that in olive, a homomorphic di-allelic self-incompatibility system is governed by a hemizygous *S*-haplotype present in only one of the two incompatibility groups. Beyond its economic importance for olive production, the identification of the *S*-locus in olive is key to understand the unique homomorphic DSI system present in Oleaceae. Hemizygoty prevents recombination and explains *S* dominance in *Ss* individuals. If the elucidation of its genetic architecture brings insights on the maintenance of the DSI, its emergence is still unclear. The parallel with the architecture of distyly supergene (Li *et al.*, 2016; Shore *et al.*, 2019; Gutiérrez-Valencia *et al.*, 2022) reinforces the question of its homology with the ancestral homomorphic DSI still present in diploid lineages of Oleaceae. Our results open the way for comparative works between olive and distylous Oleaceae species.

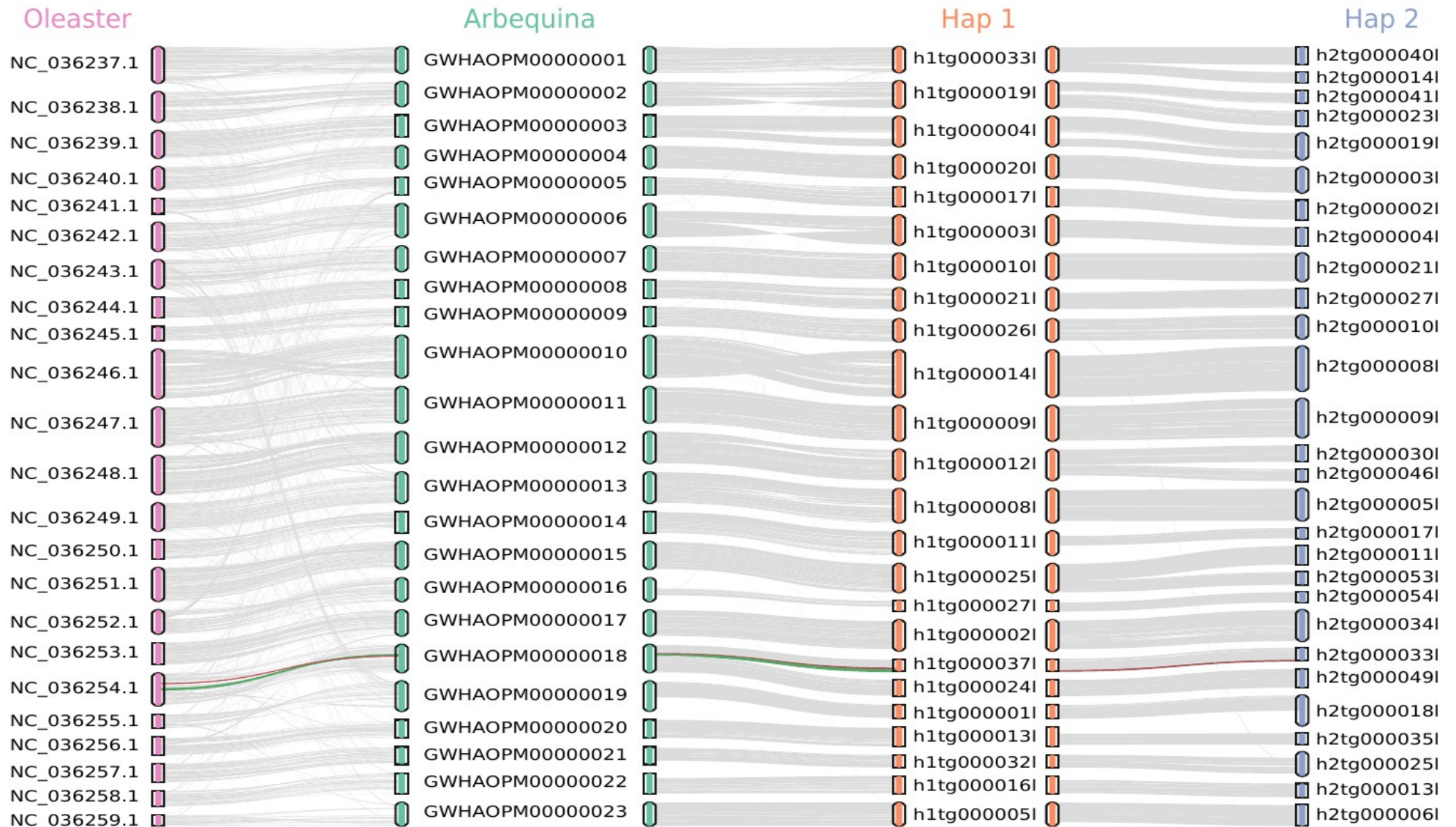


Figure 1. Macrosynteny between four olive genome assemblies. Lines connect syntenic blocks in different assemblies. Red lines correspond to the DSI locus location according to Mariotti *et al.* (2020), green lines to the newly identified SI-linked region, which has no homolog in hap2 assembly. Only anchored and major scaffolds are shown for each assembly. Hap1/Hap2 = partially-phased haplotype-resolved assemblies of *O. e. laperrinei*.

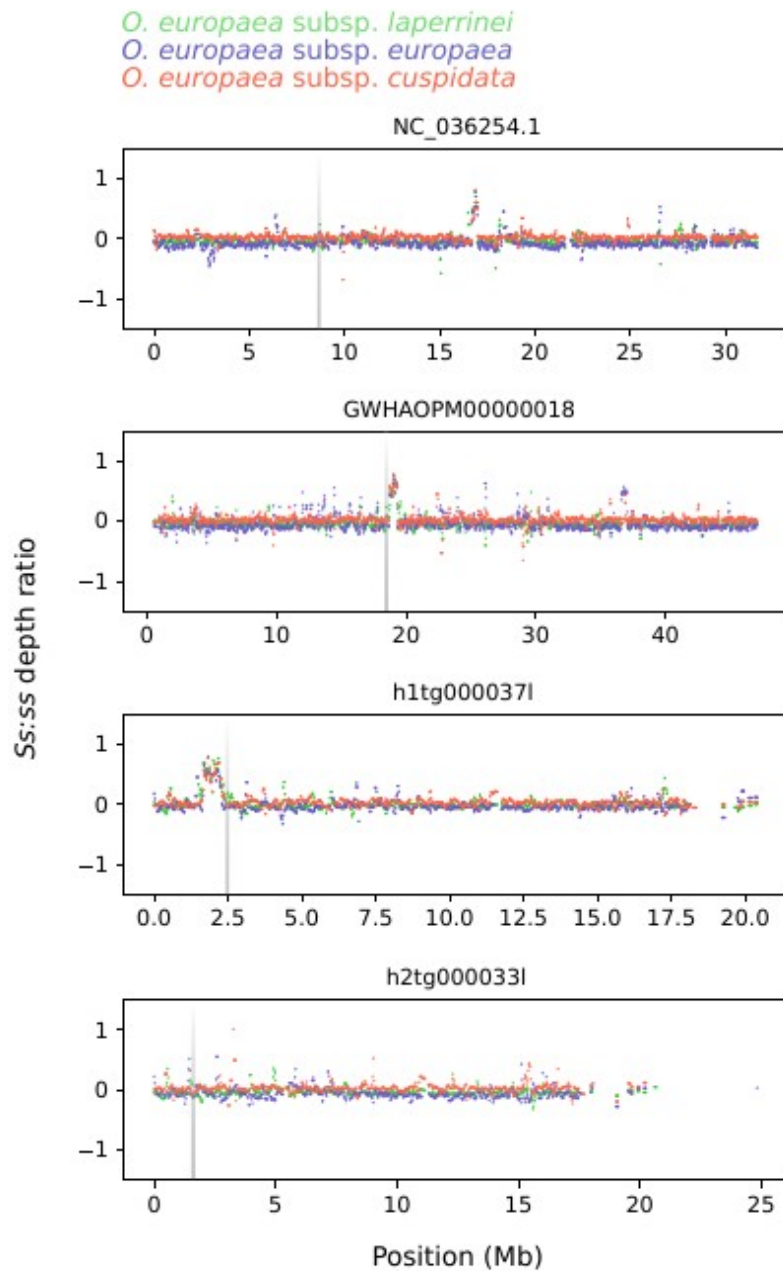


Figure 2. Variation in the ratio of depth of coverage between incompatibility groups in three olive subspecies along chromosome 18 in the different olive genome assemblies. From top to bottom, oleaster genome, cultivated olive genome, ‘haplotype 1’ and ‘haplotype 2’ of Laperrine’s olive. We plotted $\log_2[(Ss+1):(ss+1)]$ for sliding-windows of 100 kb with 25-kb steps. Values in each of the three subspecies investigated with RAD-seq are figured in a different color. The grey highlights represent the position of the previously identified SI-linked region (Mariotti *et al.*, 2020).

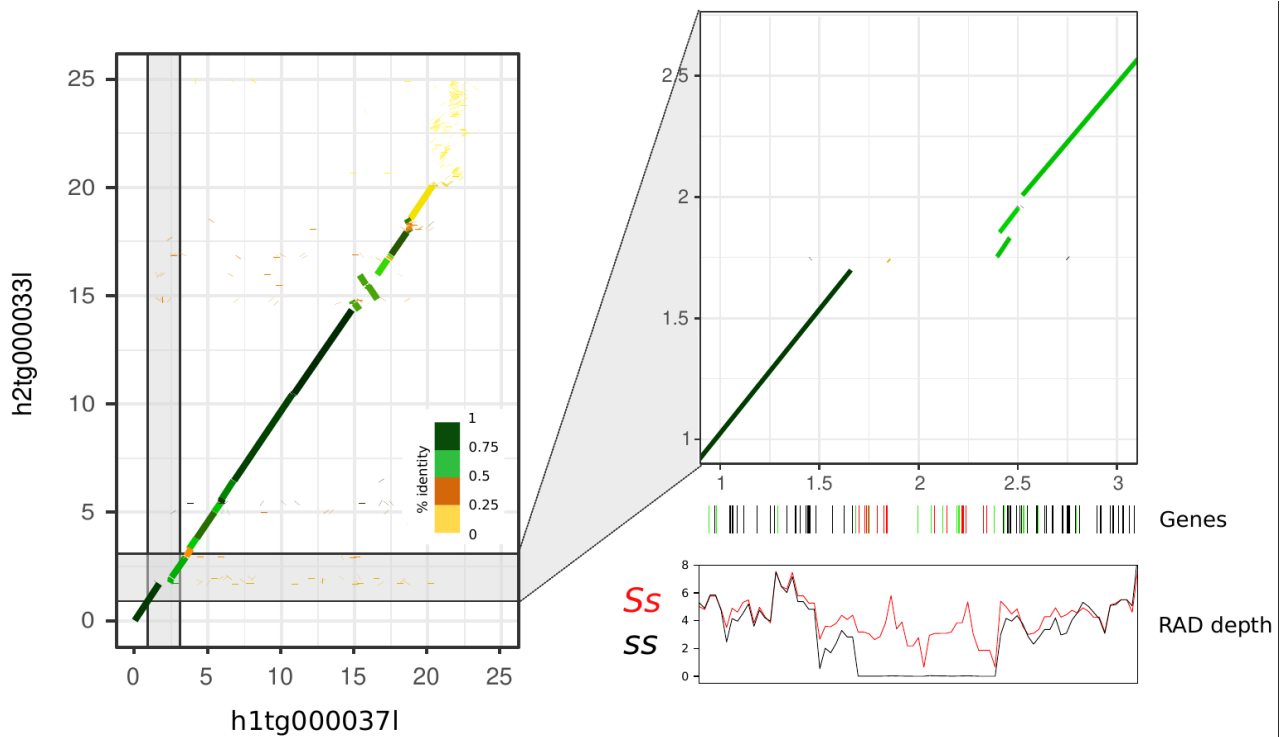


Figure 3. Pairwise synteny between the two haploid sequences of chromosome 18 from our haplotype-resolved assembly of *O. e. laperrinei*. The left dotplot presents pairwise identity over the whole scaffold, and a zoom of the region identified with coverage ratio analysis is given on the right. The gene content and median RAD depth of coverage in each incompatibility group (*Ss* in red, *ss* in black) are plotted along ‘haplotype’ 1 axis. Transposable elements are colored in green and genes within the region specific to ‘haplotype 1’ in red.

Author contributions

G.B., P-A.C., and P.R. designed the study with inputs from J.C. G.B. and P-O.C. managed olive collections. G.B. phenotyped plant material, established the sampling. G.B., S.M. did the lab work. S.V. and W.M. managed the sequencing and assembly of the Laperrine's olive genome. P.R. analyzed the genomic data and wrote the paper, with inputs from G.B. and P-A.C.

Acknowledgement

This work has received support from the grant GENRES (Occitanie-France Olive). P.R., J.C., and G.B. are members of the EDB laboratory supported by the excellence projects Labex CEBA (ANR-10-LABX-25-01) and Labex TULIP (ANR-10-LABX-0041), managed by the French ANR. P-O.C., and G.B. are also supported by the European Union's Horizon 2020 project GEN4OLIVE (H2020-SFS-2020-1; G.A. No. 101000427). The olive collection is managed by the Platform "Terrains d'Expériences" of the LabEx CeMEB (ANR-10-LABX-04-01). We thank Thierry Mathieu, Pauline Durbin (CEFE), H el ene Lasserre (France Olive), and Edy Spagnol for their help in collection management. We are grateful to the Genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, doi:10.15454/1.5572369328961167E12) for providing computing and storage resources.

References

- Besnard G, Cheptou PO, Debbaoui M, Lafont P, Hugueny B, Dupin J, Baali-Cherif D. 2020. Paternity tests support a diallelic self-incompatibility system in a wild olive (*Olea europaea* subsp. *laperrinei*, Oleaceae). *Ecol. Evol.* 10: 1876–1888.
- Br una T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics Bioinformatics* 3: iqaa108.
- Carr e A, Gallina S, Santoni S, Vernet P, God e C, Castric V, Saumitou-Laprade P. 2021. Genetic mapping of sex and self-incompatibility determinants in the androdioecious plant *Phillyrea angustifolia*. *Peer Comm. J.* 1: e15.
- de Cauwer I, Vernet P, Billiard S, God e C, Bourceaux A, Ponitzki C, Saumitou-Laprade P. 2021. Widespread coexistence of self-compatible and self-incompatible phenotypes in a diallelic self-incompatibility system in *Ligustrum vulgare* (Oleaceae). *Heredity* 127: 384–392.
- Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* 18: 170–175.
- Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* 10: giab008.
- Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA. 2011. SNP discovery and genotyping for evolutionary genetics using RAD sequencing BT - Molecular Methods for Evolutionary Genetics. In: Orgogozo V, Rockman M V, editors. Totowa, NJ: Humana Press. p. 157–178.
- Fujii S, Kubo KI, Takayama S. 2016. Non-self- and self-recognition models in plant self-incompatibility. *Nat. Plants* 2: 16130.
- Girgis HZ. 2015. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinformatics* 16: 227.
- Guti errez-Valencia J, Fracassetti M, Berdan EL, Bunikis I, Soler L, Dainat J, Kutschera VE, Losvik A, D esamor e A, Hughes PW, et al. 2022. Genomic analyses of the distyly supergene reveal convergent evolution at the molecular level. *Curr. Biol.* 32:10.1016/j.cub.2022.08.042.

- Hunter JD. 2007. Matplotlib: A 2D Graphics Environmen. *Computing in Science & Engineering*. 3:90-95.
- Huu CN, Plaschil S, Himmelbach A, Kappel C, Lenhard M. 2022. Female self-incompatibility type in heterostylous *Primula* is determined by the brassinosteroid-inactivating cytochrome P450 CYP734A50. *Curr. Biol.* 32:671-676.e5.
- Igic B, Lande R, Kohn JR. 2008. Loss of self-incompatibility and its evolutionary consequences. *Int. J. Plant Sci.* 169: 93–104.
- Jombart T, Ahmed I. 2011. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 27: 3070–3071.
- Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, *et al.* 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30: 1236–1240.
- Julca I, Marcet-Houben M, Vargas P, Gabaldón T. 2018. Phylogenomics of the olive tree (*Olea europaea*) reveals the relative contribution of ancient allo- and autopolyploidization events. *BMC Biol.* 16: 15.
- Kappel C, Huu CN, Lenhard M. 2017. A short story gets longer: Recent insights into the molecular basis of heterostyly. *J. Exp. Bot.* 68:5719–5730.
- Kamau E, Charlesworth D. 2005. Balancing selection and low recombination affect diversity near the self-incompatibility loci of the plant *Arabidopsis lyrata*. *Curr. Biol.* 15: 1773–1778.
- Knaus BJ, Grünwald NJ. 2017. vcfr: a package to manipulate and visualize variant call format data in R. *Mol. Ecol. Resour.* 17: 44–53.
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9: 357–359.
- Lawrence, M. J. 2000 Population genetics of the homomorphic SI polymorphism in flowering plants. *Ann. Bot.* 85:221–226.
- Li J, Cocker JM, Wright J, Webster MA, McMullan M, Dyer S, Swarbreck D, Caccamo M, Van Oosterhout C, Gilmartin PM. 2016. Genetic architecture and evolution of the *S*-locus supergene in *Primula vulgaris*. *Nat. Plants* 2: 16188.
- Li, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34:3094-3100.
- Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. 2017. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* 33: 574–576.
- Mariotti R, Fornasiero A, Mousavi S, Cultrera NGM, Brizioli F, Pandolfi S, Passeri V, Rossi M, Magris G, Scalabrin S, *et al.* 2020. Genetic mapping of the incompatibility locus in olive and development of a linked sequence-tagged site marker. *Front. Plant Sci.* 10: 1760.
- Matzke CM, Shore JS, Neff MM, McCubbin AG. 2020. The *Turnera* style *S*-locus gene *tsbahd* possesses brassinosteroid-inactivating activity when expressed in *Arabidopsis thaliana*. *Plants* 9:1–13.
- Olofsson JK, Cantera I, Van de Paer C, Hong-Wa C, Zedane L, Dunning LT, Alberti A, Christin PA, Besnard G. 2019. Phylogenomics using low-depth whole genome sequencing: A case study with the olive tribe. *Mol. Ecol. Resour.* 19: 877–892.

- Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, *et al.* 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* 20: 275.
- Palmer DH, Rogers TF, Dean R, Wright AE. 2019. How to identify sex chromosomes and their turnover. *Mol. Ecol.* 28: 4709–4724.
- Qiu S, Bergero R, Guirao-Rico S, Campos JL, Cezard T, Gharbi K, Charlesworth D. 2016. RAD mapping reveals an evolving, polymorphic and fuzzy boundary of a plant pseudoautosomal region. *Mol. Ecol.* 25: 414–430.
- Rao G, Zhang J, Liu X, Lin C, Xin H, Xue L, Wang C. 2021. De novo assembly of a new *Olea europaea* genome accession using nanopore sequencing. *Hortic. Res.* 8: 64.
- Rochette NC, Rivera-Colón AG, Catchen JM. 2019. Stacks 2: Analytical methods for paired end sequencing improve RADseq-based population genomics. *Mol. Ecol.* 28: 4737–4754.
- Saumitou-Laprade P, Vernet P, Vassiliadis C, Hoareau Y, De Magny G, Dommée B, Lepart J. 2010. A self-incompatibility system explains high male frequencies in an androdioecious plant. *Science* 327: 1648–1650.
- Saumitou-Laprade P, Vernet P, Vekemans X, Billiard S, Gallina S, Essalouh L, Mhaïa A, Moukhli A, El Bakkali A, Barcaccia G, *et al.* 2017. Elucidation of the genetic architecture of self-incompatibility in olive: Evolutionary consequences and perspectives for orchard management. *Evol. Appl.* 10: 867–880.
- Shen W, Le S, Li Y, Hu F. 2016. SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q File manipulation. *PLoS One* 11: e0163962.
- Shore JS, Hamam HJ, Chafe PDJ, Labonne JDJ, Henning PM, McCubbin AG. 2019. The long and short of the *S*-locus in *Turnera* (Passifloraceae). *New Phytol.* 224: 1316–1329.
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31: 3210–3212.
- Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. 2008. Synteny and collinearity in plant genomes. *Science* 320: 486–488.
- Taylor H. 1945. Cyto-taxonomy and phylogeny of the Oleaceae. *Brittonia* 5:337–367.
- Unver T, Wu Z, Sterck L, Turktas M, Lohaus R, Li Z, Yang M, He L, Deng T, Escalante FJ, *et al.* 2017. Genome of wild olive and the evolution of oil biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* 114: E9413–E9422.
- Uyenoyama MK. 2005. Evolution under tight linkage to mating type. *New Phytol.* 165:63–70.
- Vernet P, Lepercq P, Billiard S, Bourceaux A, Lepart J, Dommée B, Saumitou-Laprade P. 2016. Evidence for the long-term maintenance of a rare self-incompatibility system in Oleaceae. *New Phytol.* 210: 1408–1417.
- Zdobnov EM, Kuznetsov D, Tegenfeldt F, Manni M, Berkeley M, Kriventseva E V. 2021. OrthoDB in 2020: evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* 49: D389–D393.
- Zhang C, Zhang T, Luebert F, Xiang Y, Huang C-H, Hu Y, Rees M, Frohlich MW, Qi J, Weigend M, *et al.* 2020. Asterid phylogenomics/phylotranscriptomics uncover morphological evolutionary histories and support phylogenetic placement for numerous whole-genome duplications. *Mol. Biol. Evol.* 37: 3188–3210.

Appendix for Chapter 3

Supporting information includes the following items:

Figure S1. k-mer spectra of the two Laperrine's haplotype assemblies.

Figure S2. PCA plot of sequenced olive individuals based on bi-allelic nuclear SNPs called on four different genome assemblies.

Figure S3. F_{ST} indices between incompatibility groups along each chromosomal scaffold in four olive genome assemblies.

Figure S4. Variation in coverage ratio between incompatibility groups along each chromosome scaffold in four olive genome assemblies.

Figure S5. Gene density along chromosome 18 in the 'Arbequina' genome.

Table S1. List of plant material analyzed with RAD-sequencing.

Table S2. RAD-sequencing and mapping statistics on the oleaster genome.

Table S3. Predicted functions of *S*-locus genes and correspondence between genome annotation.

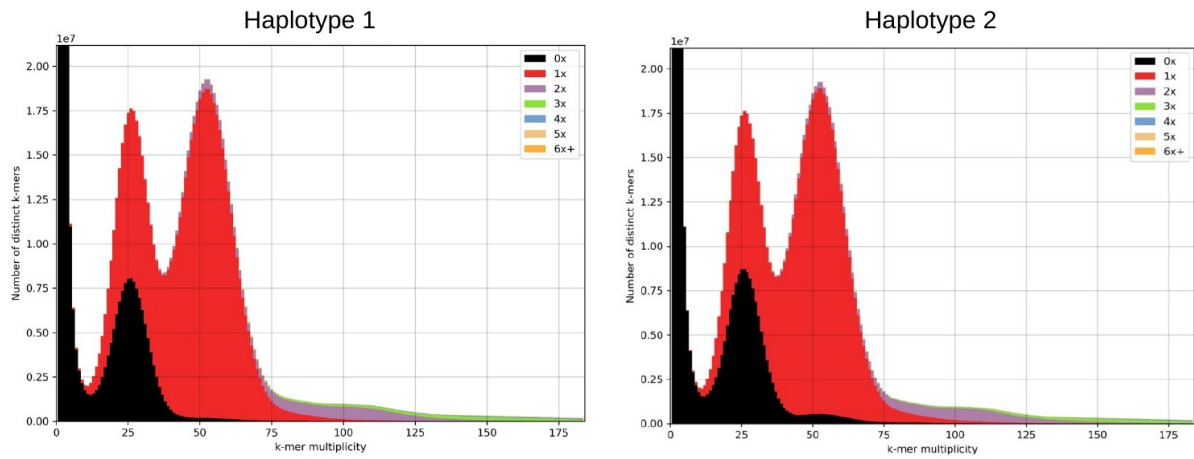


Figure S1. *k*-mer spectra of the two Laperrine's haplotype assemblies. Red curves correspond to *k*-mer present in one copy in the assembly (the two peaks at 25 \times and 50 \times reflects the high heterozygosity of the genome). Black curve represents *k*-mer in reads that were not included in the assembly, either erroneous (at the extreme boundary of the plot) or because of allele separation (peak half the size of the heterozygous peak). Other colors indicate higher *k*-mer multiplicity probably due to ancient whole-genome duplications.

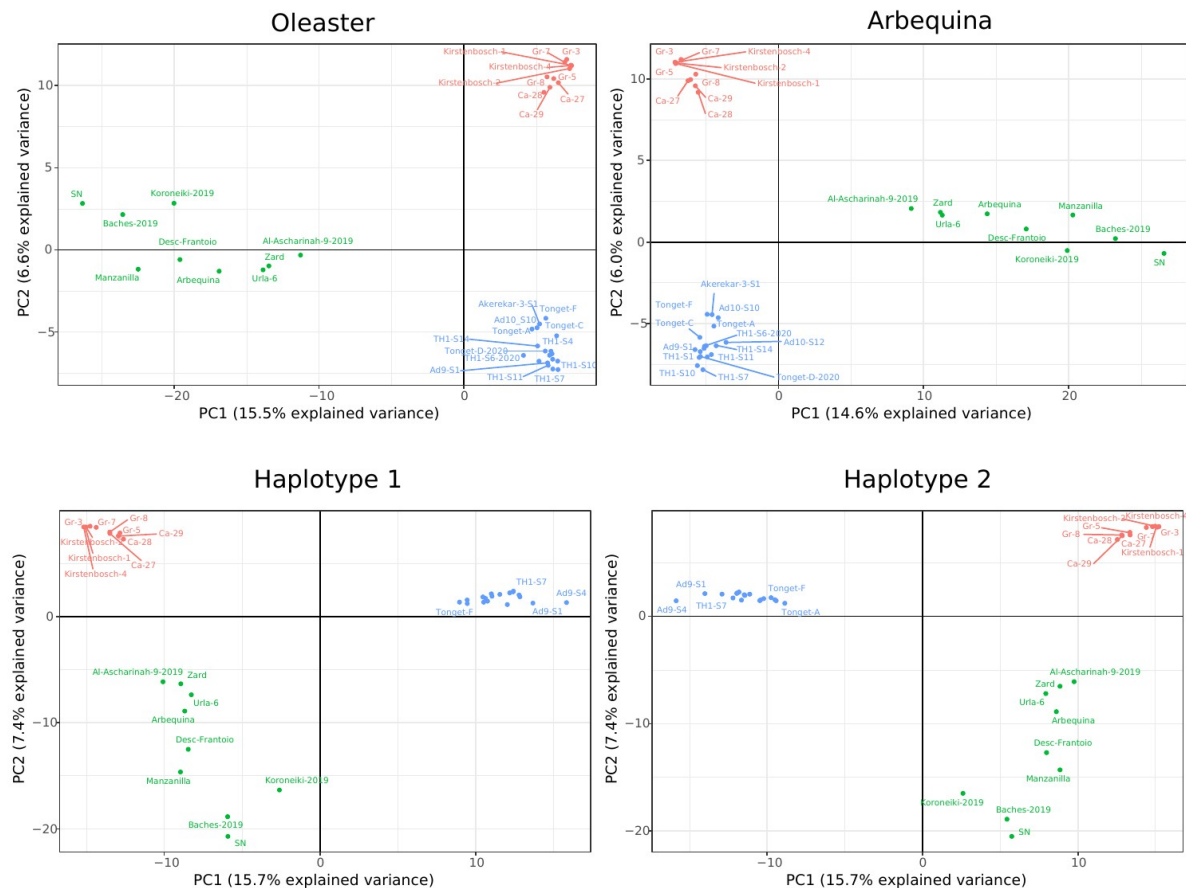


Figure S2. PCA component analysis plot of sequenced olive individuals based on bi-allelic nuclear SNPs called on four different genome assemblies. Each subspecies is figured in a different color. Blue=*Olea e. europaea*, green=*O. e. laperrinei*, orange=*O. e. cuspidata*. Number of SNPs used for each genome: Oleaster 43,525 / Arbequina 41,508 / Haplotype 1 48,623 / Haplotype 2 48,623.

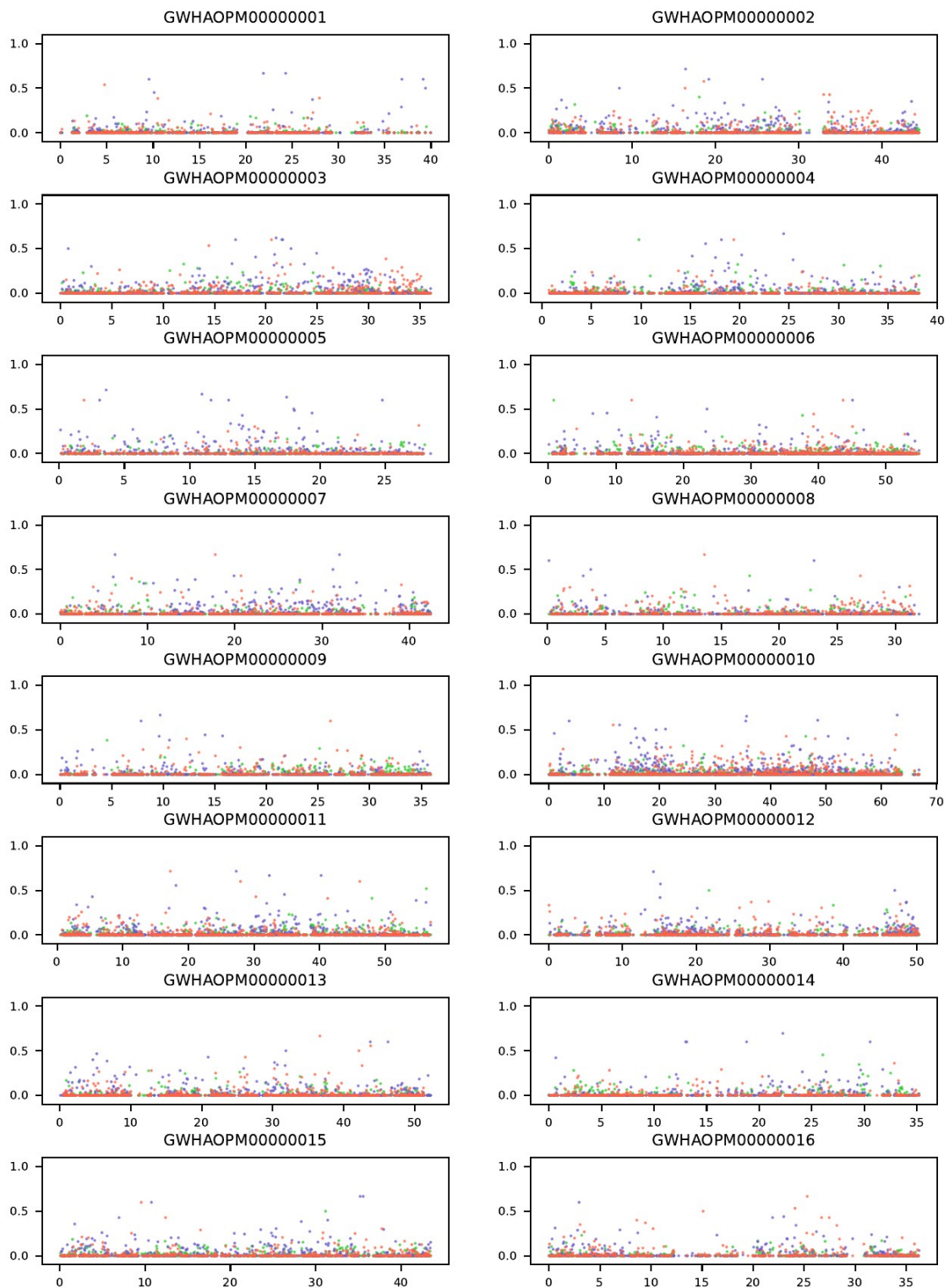


Figure S3. F_{ST} indices between incompatibility groups along each chromosomal scaffold in four olive genome assemblies. F_{ST} values were calculated for non-overlapping windows of 50 kb. Values in each of the three subspecies investigated with RAD-seq are figured in a different color. Blue = *Olea e. europaea*, green = *O. e. laperrinei*, orange = *O. e. cuspidata*. The remaining chromosomes are available at : https://github.com/praimondeau/olive_SI

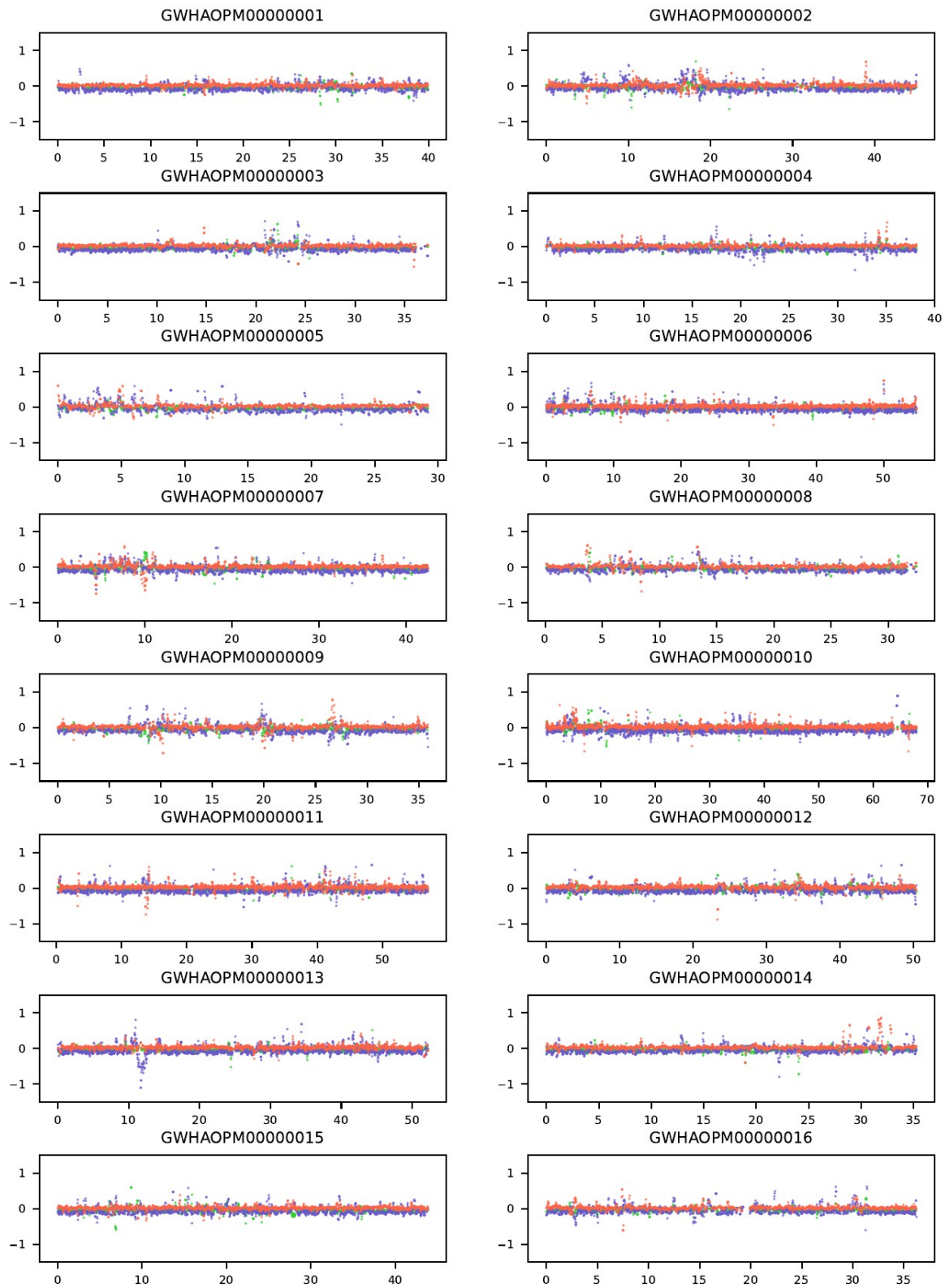


Figure S4. Variation in coverage ratio between incompatibility groups along each chromosomal scaffold in four olive genome assemblies. We plotted $\log_2[(Ss+1):(ss+1)]$ for sliding-windows of 100 kb with 25-kb steps. Values in each of the three subspecies investigated with RAD-seq are figured in a different color. Blue = *Olea e. europaea*, green = *O. e. laperrinei*, orange = *O. e. cuspidata*. The remaining chromosomes are available at : https://github.com/praimondeau/olive_SI

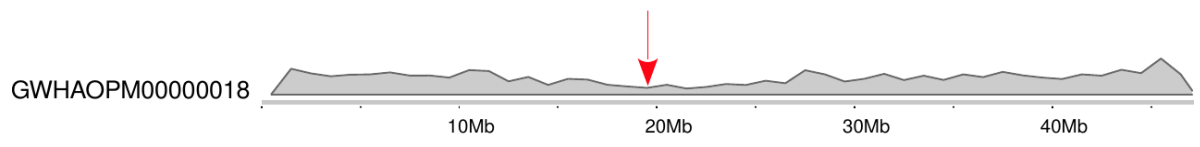


Figure S5. Gene density along chromosome 18 in the 'Arbequina' genome. The red arrow indicates the *S*-locus location.

Table S1. List of plant material analyzed with RAD-sequencing.

Subspecies	Accession name	CC ¹ group	Subspecies	Accession name	CC group
<i>europaea</i>	L4R10 ('Zard')	G2	<i>laperrinei</i>	Adjelella 9_S1	G2
	L4R11 ('Manzanilla')	G2		Adjelella 9_S4*	G1
	L4R13 ('desc. Frantoio')	G1		Adjelella 10_S9	G2
	L4R14 ('Koroneiki')	G1		Adjelella 10_S10	G1
	L4R15 ('Arbequina')	G1		Adjelella 10_S11	G2
	L4R17 ('Baches')	G2		Adjelella 10_S12	G1
	L4R19 ('SN')	G1		Tin-Hamor S1	G2
	L4R24 ('Urla 6')	G2		Tin-Hamor S4	G1
	L4R25 ('Al Ascharinah 9')	G2		Tin-Hamor S6	G2
				Tin-Hamor S7	G2
<i>cuspidata</i>	Gr 3	G1		Tin-Hamor S10	G1
	Gr 5	G1		Tin-Hamor S11	G1
	Gr 7	G1		Tin-Hamor S14	G1
	Gr 8	G2		Akerakar 3_S1	G1
	Ca 27	G2		Tonget A	G2
	Ca 28	G1		Tonget C	G1
	Ca 29	G2		Tonget F	G2
	Kirst 1	G2		Tonget D	G1
	Kirst 2	G2			
	Kirst 4	G1			

¹CC group were determined in Besnard et al. (2020)

Table S2. RAD-sequencing and mapping statistics on the oleaster genome.

Sample	Raw reads	Clean reads	Mapping rate (%)	Sample	Raw reads	Clean reads	Mapping rate (%)
Adjelella 10_S10	10048146	10011485	82.0	Kirst_1	21128268	21058689	82.2
Adjelella 10_S11	16924894	16840808	80.0	Kirst_2	18835768	18732614	80.7
Adjelella 10_S12	11666302	11583285	80.1	Kirst_4	21805072	21708485	82.0
Adjelella 10_S9	14910068	14866348	80.4	Koroneiki (L4R14)	9141566	9092291	80.5
Adjelella 9_S1	25275302	25107646	80.7	Manzanilla (L4R11)	24268672	24114103	82.1
Adjelella 9_S4	27142444	27027369	82.1	SN (L4R19)	20343488	20287882	80.2
Akerekar 3_S1	11921620	11885430	82.1	Tin-Hamor1_S1	18233404	18162892	80.2
Al_Ascharinah_9 (L4R25)	17485822	17415930	83.0	Tin-Hamor1_S10	18628334	18553433	81.5
Arbequina (L4R14)	13088002	13038573	80.1	Tin-Hamor1_S11	15629066	15572008	80.0
Baches (L4R17)	18660040	18573078	83.1	Tin-Hamor1_S14	12274968	12228457	80.4
Ca 27	14619204	14555327	79.9	Tin-Hamor1_S4	16914898	16848388	80.3
Ca 28	12719166	12624495	79.1	Tin-Hamor1_S6	15823176	15751436	78.4
Ca 29	14819554	14688804	76.7	Tin-Hamor1_S7	20462340	20304506	73.9
Desc Frantoio (L4R15)	16395184	16354789	82.9	Tonget A	12489254	12410717	77.9
Gr 3	20845246	20740413	79.8	Tonget C	18353244	18303637	78.0
Gr 5	15587608	15520209	79.2	Tonget D	18371078	18306346	79.7
Gr 7	19708102	19625567	81.2	Tonget F	14995700	14944628	79.0
Gr 8	14484088	14397980	80.3	Urla 6 (L4R24)	14229824	14174075	81.7
				Zard (L4R10)	12930964	12879725	84.4

Table S3. Predicted function of *S*-locus genes and correspondence between genome annotation.

Gene ID in 'haplotype 1'	Gene ID in 'Arbequina'	Gene ID in oleaster	Prediction about function
g33207.t1	-	-	-
g33208.t1	-	-	-
g33210.t1	GWHTAOPM038200	LOC111366801	-
g33212.t1	NA ¹	-	-
g33216.t1	-	-	TE-related
g33217.t2	-	-	-
g33218.t1	-	-	-
g33219.t1	-	-	IPP transferase
g33222.t1	-	-	TE-related
g33223.t1	GWHPAOPM038191	LOC111379898	BES1/BZR1 plant transcription factor homolog
g33228.t1	-	-	-
g33230.t1	-	-	-
g33231.t1	GWHPAOPM038189	LOC111392691	Aminotransferase-like, PLP-dependent enzymes
g33232.t1	-	-	TE-related
g33233.t1	GWHPAOPM038188	LOC111392689	Gibberellin 2-beta-dioxygenase 2
g33235.t1	-	-	TE-related
g33236.t2	-	-	Exostosin heparan sulfate glycosyltransferase homolog

¹Corresponding sequence is within a deletion in this cultivar.

Chapter 4

Transcriptomics of the development of distyly in jasmine

Raimondeau P., Valière S., Besnard G.

Abstract

In distylous lineages, such as jasmine (Oleaceae), individuals are shared between two types of flowers differing in stamens and style heights. This morphological variation favors cross-pollination and evolved repetitively in flowering plants. Characterising the molecular basis underlying this floral polymorphism in distinct lineages is crucial to understand the evolutionary pathways to distyly. In jasmine, the genetics of distyly remains completely unexplored. We undertook the comparison of expression profiles between the two floral morphs of the Mediterranean jasmine (*C. fruticans*) before anthesis. Transcriptomic changes were discreet. Most of the differentially expressed genes were linked to the presence of a virus in the short-styled individuals that may indicate greater sensibility in this morph. We still determined a short list of differentially expressed genes between morphs that will need further investigation.

Key-words: heterostyly, RNA-seq, jasmine, differential expression, virus

Introduction

Heterostyly is a genetically determined floral polymorphism in which individuals within a species vary in the positions of their sexual organs. In distylous lineages, individuals are shared between two types of flowers differing in the heights at which the stamens and style are positioned. Long-styled individuals (L-morph) present low anthers while short-styled individuals (S-morph) have higher anthers. Heterostyly evolved independently multiple time and has been identified in 28 families (Barrett, 2019). This is a remarkable example of convergent evolution to promote outbreeding. Indeed, this morphological variation favors the deposition of pollen grains from each morph onto different regions of pollinators' bodies and consequently on pistils from the opposite morph (Darwin, 1877). Heterostyly is usually associated with a heteromorphic self-incompatibility system that further reduces the risk of inbreeding (Barrett, 1998).

Distyly and its genetic control have been intriguing biologists for more than a century, most efforts being focused on the genus *Primula* (Gilmartin, 2015). It was early established to be controlled by a single locus, with a dominant *S* allele, present only in *S*-morph individuals and a recessive *s* allele, in homozygous state in L-morph plants (Bateson and Gregory, 1905). This model was later refined to a *S*-locus in the form of a supergene composed of at least three tightly linked genes controlling the different features of heterostyly: style length, anther length, and pollen (Ernst, 1936), but also the genetic determinants of heteromorphic self-incompatibility. In *Primula*, it was recently shown that the dominant *S* haplotype is a hemizygous 278-kb region exclusive to *S*-morph individuals (Li *et al.*, 2016). The hemizygous region contains five genes, one controls both the

position of the style (Huu *et al.*, 2016) and female compatibility (Huu *et al.*, 2022), and another determines anther position (Huu *et al.*, 2020). The component responsible for male compatibility is still unknown. Other heterostylous systems have been partially characterized in *Turnera* (Shore *et al.*, 2019; Matzke *et al.*, 2020; Matzke *et al.*, 2021), *Linum* (Ushijima *et al.*, 2012; Gutiérrez-Valencia *et al.*, 2022) and *Fagopyrum* (Yasui *et al.*, 2012). The multiple independent evolutions of heterostyly make deciphering the mechanisms of each system particularly interesting to understand the common features underlying their emergence. To date, two theoretical models for the development of heterostyly have been proposed (Charlesworth and Charlesworth, 1979; Lloyd and Webb, 1992) with some experimental evidence supporting either of them (reviewed in Barrett, 2019).

In Oleaceae, distyly has been traditionally considered as the ancestral state (Taylor, 1945) and was reported in several lineages of Myxopyreae (Kiew, 1984), Forsythieae (Sampson, 1971; Ryu *et al.*, 1976), Schreberinae (Perrier de la Bâthie, 1951; Verdoorn, 1956), and Jasmineae (Domme *et al.*, 1992; Olesen *et al.*, 2003; Ganguly and Barua, 2020). The sister family Carlemaniaceae (Zhang *et al.*, 2020) also includes distylous species (Tange, 1998). The bulk of studies on the biology of distylous and herkogamous jasmines (Domme *et al.*, 1992; Guitián *et al.*, 1998; Thompson and Domme 2000; Olesen *et al.*, 2003; Ganguly and Barua, 2020), particularly focused on the Mediterranean species *Chrysojasminum fruticans* (L.) Banfi. Difference between the two morphs in morphology, phenology, reproductive success and fruit set have thus been investigated, but the molecular basis of distyly in Oleaceae remains completely unexplored.

In this work, we attempted the first characterization of the genetic basis of heterostyly in a distylous Oleaceae, the Mediterranean jasmine (*C. fruticans*). Using floral bud tissues from both morphs, we generated expression data at two time points during floral development and provide a reference transcriptome for the species. We then performed differential expression analysis between long-styled and short-styled individuals to try and identify candidate genes involved in morph differentiation.

Materials and Methods

Plant material and RNA isolation

Plants were collected (on April 29, 2021) at the common garden of CEFE in Montpellier, France. *Chrysojasminum fruticans* buds were collected at two developmental stages at which floral morphology is thought to diverge between mating types (bud length 0.5 cm and 1 cm approximately). At the first stage (0.5 cm), buds are green. About three days later (1 cm), buds are yellow and about to open the next day or the day after depending on environmental conditions. Five

distinct individuals were sampled for each morphotype (as determined based on the observation of more advanced flowers on the same plant; for each individual the same phenotype was also recorded in 2022). In addition, we similarly collected buds at the same stages on a L-morph individual of the sister Canarian species *Chrysojasminum odoratissimum* (L.) Banfi (see nuclear phylogeny in Chapter 2, Supplementary Notes) and on an individual of the non-heterostylous and self-compatible species *Chrysojasminum bignoniaceum* (both maintained in our collection at EDB). Samples were placed in liquid nitrogen on the field right after their collection, transported on dry ice and kept at -80°C. RNA extractions were performed with the Qiagen RNeasy Plant Mini Kit following manufacturer's instructions. Extracts were then purified with Promega RQ1 RNase Free DNase to limit DNA contamination. Quality controls and library preparation were performed at the GenoToul sequencing platform (Toulouse, France) using Illumina TruSeq Stranded mRNA preparation Kit.

Transcriptome sequencing and assembly

One SP lane of Illumina NovaSeq was used to generate 150-bp paired-end reads for 24 jasmine samples. Raw reads were cleaned and trimmed using Fastp (Chen *et al.*, 2018) with default parameters to remove any low-quality bases or adapters. Clean reads of *C. fruticans* were then pooled to produce a single transcriptome assembly. *De novo* transcriptome assembly was performed with Trinity using the strand-specific option, *in silico* normalization and a minimal *k*-mer coverage of 2. The E90N50 and the BUSCO score against eudicots_odb10 (Simão *et al.*, 2015) were examined to assess the quality of the resulting assembly. In addition, to estimate the read representation of the assembly, clean paired-end reads were mapped back on the assembled transcripts using Bowtie2 (Langmead and Salzberg, 2012). We also examine the representation of full-length protein-coding genes, by blasting the assembled transcripts against the published *Jasminum sambac* protein sequences (Xu *et al.*, 2022). Transcripts were annotated using Trinotate (Bryant *et al.*, 2017).

Differential expression analysis

For each sample, read counts were summarized from reads mapped on the complete set of transcripts using RSEM (Li and Dewey, 2011) as part of the Trinity analysis pipeline. Count Per Million (CPM) were used to normalize for library sizes, prior to explore samples relationship using Pearson's correlation coefficient and Principal Component Analysis with the function `prcomp`. Only transcripts with at least 0.5 CPM in more than three samples were further considered. Analyses of differential transcripts expression between morphs at each stage were conducted with EdgeR

(Robinson *et al.*, 2010) using TMM (Trimmed Mean of M-values) to account for differences in library composition. Differentially expressed transcripts (fold change >2 and false discovery rate < 0.05) were clustered according to Euclidian distance between expression profiles. We check for gene ontology enrichment in differentially expressed features using the package goseq (Young *et al.*, 2010) and the GO annotation from Trinotate output.

Results

Transcriptome assembly

Extraction yielded 15 to 33 ug of RNA per sample that was used to generate 24 RNA-Seq libraries for twelve *Chrysojasminum* individuals. Sequencing produced 958 million paired end 150-bp reads (between 32 and 46 million reads per sample). Between 95 and 98% of raw reads were kept after cleaning, resulting in 774 million paired-reads being used to generate the transcriptome assembly for *C. fruticans*. 242,066 transcripts were assembled with a Ex90N50 of 2,032 bp. Most of them are probably not biologically relevant and represents lowly expressed transcripts. If we consider transcripts with a minimum TPM (Transcripts Per Million transcripts) of 10, the number of transcripts is reduced to 30,150 (Figure S1). The completeness of the assembly was assessed by the high percentage of reads that mapped back to the assembly (98%, 81% concordantly) and the fact that 96.7% of eudicots BUSCO genes were detected as complete, and only 1.1% as fragmented. 12,000 transcripts were estimated to cover 90% or more of *Jasminum sambac* proteins. 183,334 transcripts contained ORF of which 116,591 were annotated successfully.

Differential expression between morphs

All samples expression profiles, regardless of the morph or stages, were strongly correlated (Pearson's correlation coefficient greater than 0.85; Figure 1A). Samples are not more related within morphs or within stages, as clearly shown by PCA results (Figure 1B). The difference between conditions is thus tenuous and resides in a small number of differentially expressed genes. We inspected transcripts differentially expressed between morph in one or both stages. When examining expression profiles (Figure 2), differentially expressed transcripts can be split in four main groups: features strongly expressed in *S*-morphs samples but not in *L*-morph (cluster A; 18 features), a second cluster with lower fold-change but similar profile (cluster B; 16 features), a third 'cluster' with a single gene highly expressed mostly in *L*-morph and finally a cluster gathering heterogeneous profiles with smaller fold-change values (cluster D; 38 features). Out of 73 genes with a fold change greater than 2, only eight show the same morph-specific pattern across both stages, 19 were only differentially expressed at stage 1, and 46 at stage 2 (Figure 3A). Sixteen

transcripts, only present in three/four short-styled individuals were annotated as of viral origin (Table S2). In fact, few additional transcripts are probably linked to the same Nepovirus (not annotated but belong to the same Trinity unigene, and have high TMM values, similarly to viral transcripts). Three transcripts annotated as “retrovirus-related polyprotein from transposon” presented a similar occurrence pattern (Figure S2). These virus-related features constitute the whole cluster A and four elements of cluster B and most of the features with the highest TMM values (Figure S2). Thus, they represent the most spectacular expression changes observed between the two morphs. Analysis of GO terms representation reflected this and all the significantly enriched GO terms are related to host and virus (Figure 3C). We checked the presence of these sequences in other sequencing data available for *C. fruticans* and other jasmine. We did not find any reads corresponding in the dataset used to assemble and annotate *Jasminum sambac* (heterostylous) genome (Xu *et al.*, 2022) or in our low-coverage (ca. 40×) whole-genome sequencing of both morph of *C. fruticans*. Among the remaining differentially expressed transcripts, we identified some potentially interesting genes. For instance, the two remaining genes up-regulated at both stages (Figure 3A,B) in *S*-morph (FBK77_1560_c0_g1_i8 and C3H46_6214_c0_g1_i47) contain a F-box and a Zing-finger domain, respectively (Table S2).

Discussion

No global expression changes between morphs during pre-anthesis

To search for genes that might underlie the observed phenotypic differences between heterostylous morphs in Jasmineae (Thompson and Dommée, 1993; Thompson and Dommée, 2000; Olesen *et al.*, 2003), we compared gene expression between floral buds of L- and *S*-morph plants in *C. fruticans*. Analyses of differential gene expressions have been successfully applied in the study of heterostyly in *Lithospermum* (Cohen, 2016), *Linum* (Ushijima *et al.*, 2012), *Fagopyrum* (Yasui *et al.*, 2012) and *Primula* (Burrows and McCubbin, 2018) where it enabled the discovery of several candidate genes involved in morph differences. We expected the morphological variation to stem from differential cell growth early in flower development (Gutián *et al.*, 1998; Dadpour *et al.*, 2011). The difference in expression profiles between morphs was discreet. A relatively small number (73) of differentially expressed features were finally identified, and none of them showed a clear-cut between the two morphs (i.e. expressed in five individuals of one morph and none of the other).

A shortlist of genes had differential pattern of expression between morphs

Among genes showing a significant differential expression between morphs, a quarter of them were related to a higher nepovirus gene expression in short-styled individuals (Figure 3). Actually, all

genes showing the greatest fold-change correspond to nepovirus (Figure 2; Figure S2). Such a higher viral abundance in the *S*-morph is unlikely to stem from contamination as samples from both morphs were processed together. It could thus reflect a greater sensibility of *S*-morph individuals to this phytovirus that would need further characterization.

The identified *C. fruticans* genes with differential morph expression may represent downstream components in the elaboration or function of the two morphs. Indeed, some variations in stigma-height or style curling were reported in the species and might arise from modifier loci (Thompson and Dommée, 2000). We here report several candidates that may be directly or indirectly involved in jasmine distyly. Among them, a F-box/kelch-repeat protein specifically expressed in *S*-morph individuals (Figure 3B) represents an interesting candidate due to its similarity to one of the *Primula S*-locus genes (Li *et al.*, 2016). Indeed, a F-box/kelch-repeat protein homolog was described as one of the five genes of the *Primula* hemizygous *S*-locus and is apparently involved in regulating cytokinin in *Arabidopsis thaliana* (Li *et al.*, 2016).

Conclusions

This work represents the first attempt to decipher the genomics of distyly in jasmine. Compared to other transcriptomic study of distyly in other models, we identified few differentially expressed genes. Two factors may have weakened the ability of our approach to detect interesting candidates involved in distyly. First, we collected whole buds. Dissecting samples to sequence only reproductive tissues would have potentially avoid diluting signal among other floral pieces that do not differ between morphs. A precise morphological description of flower development would be also be helpful to precisely target the most relevant stages for morph differentiation (e.g. arrested development of style in short-style individuals). Despite these caveats, this study provides resources for further investigation.

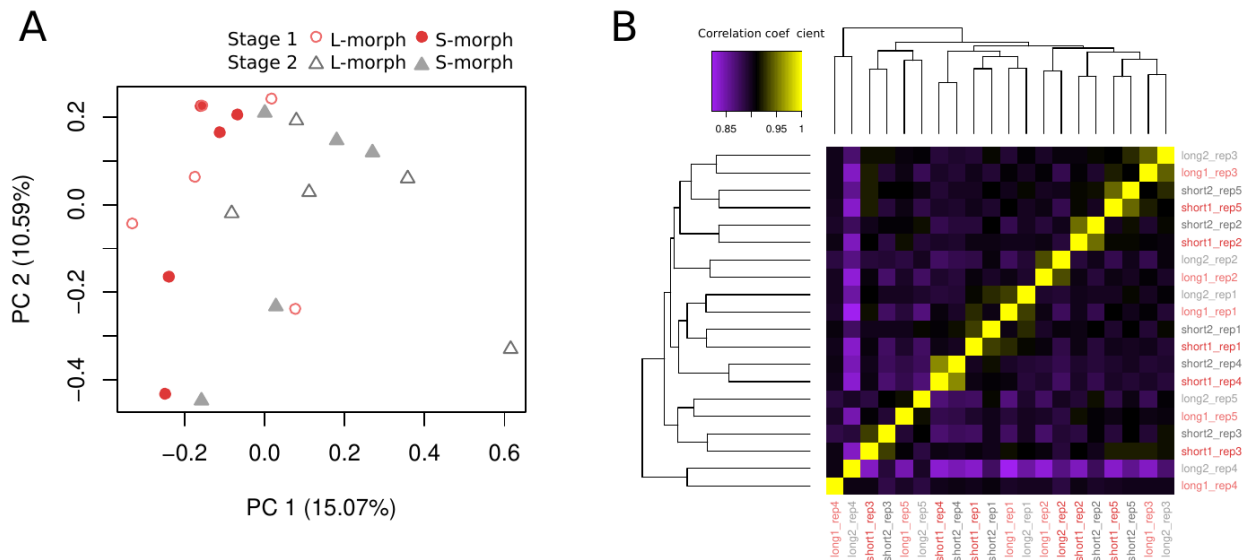


Figure 1. Relationship between samples according to expression profiles. A) PCA plot based on CPM expression values. B) Correlation matrix between samples based on CPM counts. Samples are colored according to developmental stages and darker shades/plain shapes indicate *S*-morph individuals.

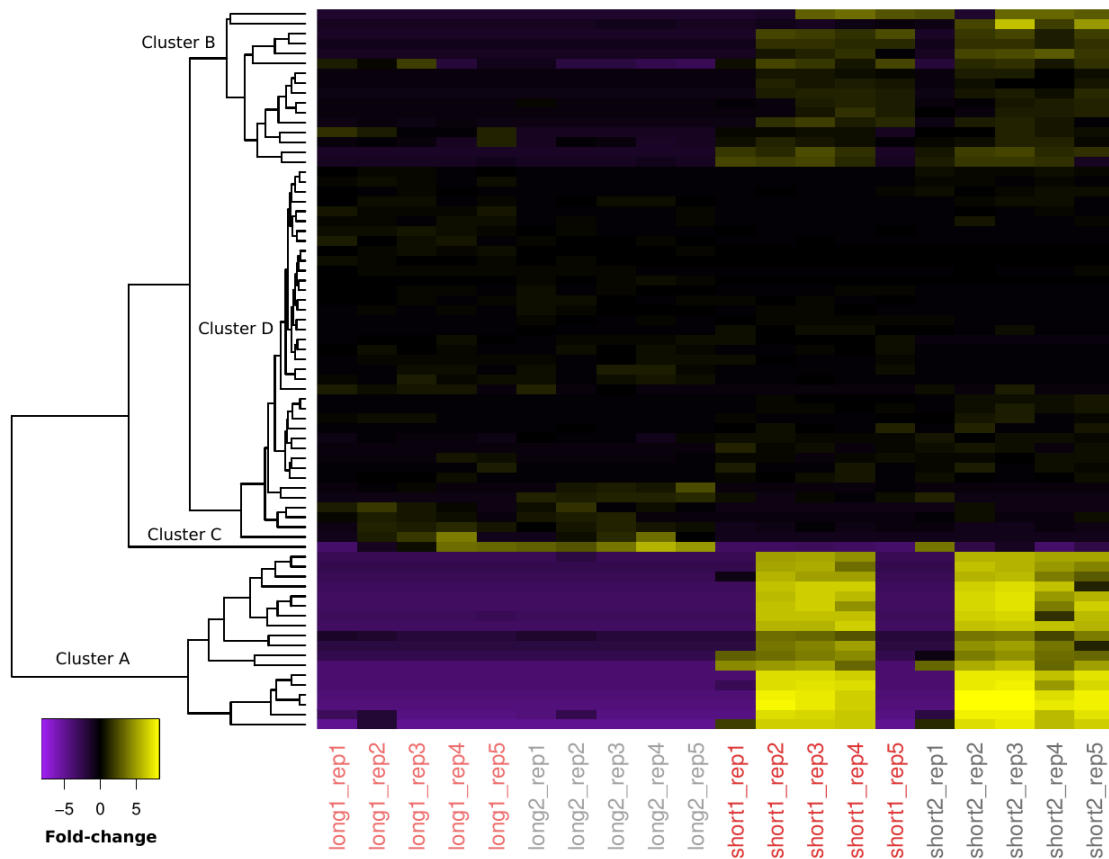
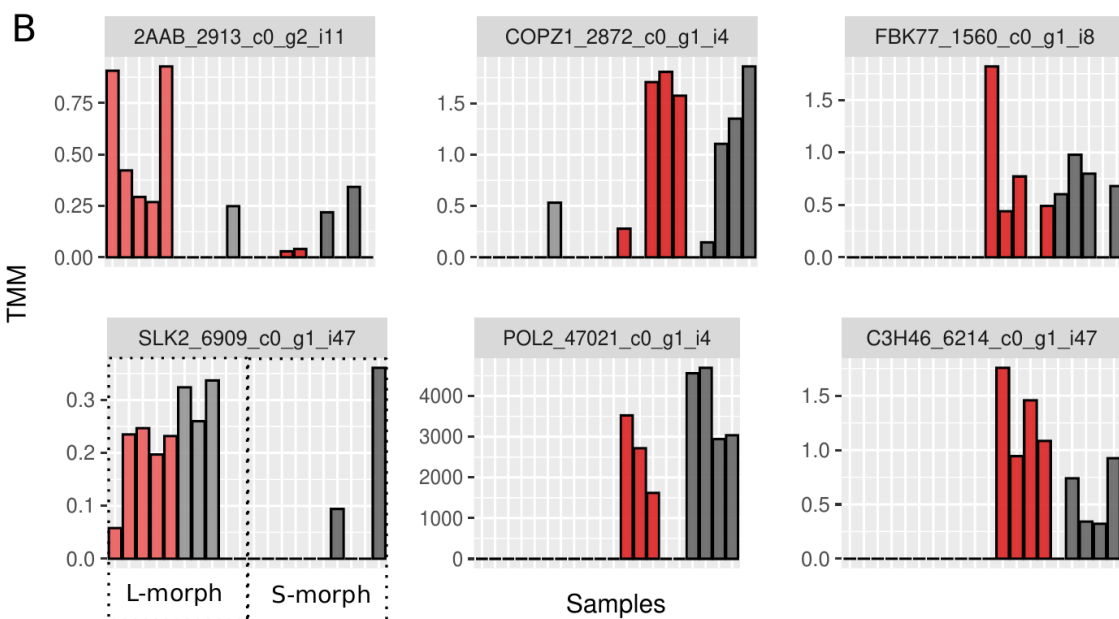
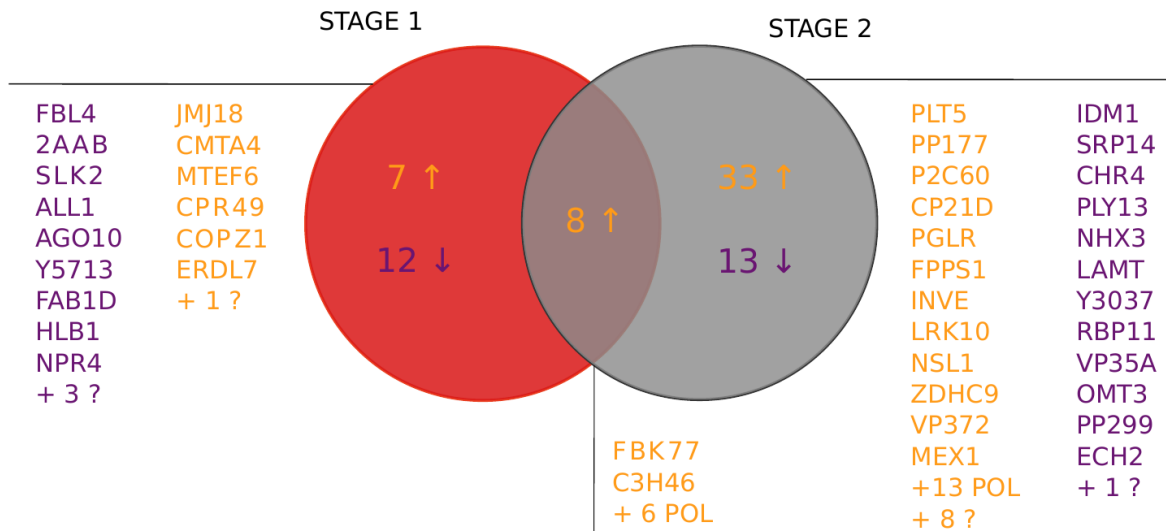


Figure 2. Clustering of transcripts according to expression profiles. Each row of the heatmap summarizes the expression of one differentially expressed transcript in each individual. Transcripts were clustered according to their Euclidian distance. Positive fold-change (yellow) indicates overexpression in a given sample compared to median expression level. Samples are colored according to developmental stages and darker shades denote *S*-morph individuals.

A Differentially expressed in S-morph



C

nucleic acid–protein covalent cross–linking
 transport of virus in multicellular host multi–organism localization
 multi–organism transport RNA–protein covalent cross–linking
 interspecies interaction between organisms multi–organism process
 viral RNA genome replication movement in host transport of virus
 viral genome replication multi–organism metabolic process
 transport of virus in host, cell to cell intercellular transport
 interaction with host multi–organism cellular process
 multi–organism intercellular transport RNA replication
 movement in host environment viral process

Figure 3. Differentially expressed genes between short and long-styled individuals. A) Venn-diagram summarizing all significant expression changes detected in *S*-morph compared to *L*-morph expression profiles at each and both stages. Yellow indicates overexpression in *S*-morph compared to *L*-morph, purple indicates the opposite. Transcripts are called by their annotated gene symbol, POL= viral polyprotein, ?=unknown function. B) Exemplar expression levels of differentially expressed transcripts in each sample. Samples are colored according to developmental stages and darker shades indicate *S*-morph individuals. C) Cloud of enriched GO-terms in the sets of differentially expressed transcripts.

Author contributions

GB and PR designed the study. GB collected the samples. GB extracted RNAs and SV prepared the libraries for sequencing. PR analyzed the data and wrote the paper, with inputs from GB.

Acknowledgement

This work was funded by a grant from TULIP-GS to PR. GB and PR are members of the Laboratoire Evolution & Diversité Biologique (EDB), part of the LABEX TULIP managed by Agence Nationale de la Recherche (ANR; no. ANR-10-LABX-0041) and were funded by GeneRes (Occitanie-France Olive). We also acknowledge an Investissement d'Avenir grant of the ANR (CEBA: ANR-10-LABX-25-01). We thank Pascal-Antoine Christin for helpful discussions, Sophie Manzi, Anne-Laure Fuchs and Elwen Le Bras for lab work assistance. We would also like to thank Thierry Mathieu and Marie-Pierre Dubois for their help organizing fieldwork and Jérôme Lluch for his advice on the sequencing strategy. We are grateful to Get-Plage and Genotoul platforms for sequencing services and providing computing resources.

References

- Barrett SCH. 1998. The evolution of mating strategies in flowering plants. *Trends Plant Sci.* 3:335–341.
- Barrett SCH. 2019. ‘A most complex marriage arrangement’: recent advances on heterostyly and unresolved questions. *New Phytol.* 224:1051–1067.
- Bateson W, Gregory RP. 1905. On the inheritance of heterostylism in *Primula*. *Proc. R. Soc. London. Ser. B, Contain. Pap. a Biol. Character* 76:581–586.
- Bryant DM, Johnson K, DiTommaso T, Tickle T, Couger MB, Payzin-Dogru D, Lee TJ, Leigh ND, Kuo T-H, Davis FG, *et al.*, 2017. A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. *Cell Rep.* 18:762–776.
- Burrows B, McCubbin A. 2018. Examination of *S*-locus regulated differential expression in *Primula vulgaris* floral development. *Plants* 7.
- Charlesworth B, Charlesworth D. 1979. The Maintenance and Breakdown of Distyly. *Am. Nat.* 114:499–513.
- Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34:i884–i890.
- Cohen JJ. 2016. De novo sequencing and comparative transcriptomics of floral development of the distylous species *Lithospermum multiflorum*. *Front. Plant Sci.* 7.
- Dadpour MR, Naghiloo S, Neycharan SF. 2011. Inflorescence and floral ontogeny in *Jasminum fruticans* (Oleaceae). *Aust. J. Bot.* 59:498–506.
- Darwin CR. 1877. *The different forms of flowers on plants of the same species*. London, UK: John Murray.
- Domme B, Thompson JD, Cristini F. 1992. Distylie chez *Jasminum fruticans* L.: hypothese de la pollinisation optimale basée sur les variations de l'écologie intraflorale. *Bull. - Soc. Bot. Fr. Lettres Bot.* 139:223–234.
- Ernst A. 1936. Heterostylie-Forschung: Versuche zur genetischen Analyse eines Organisations- und ‘Anpassungs’ merkmals. *Zeitschrift für Induktive Abstammungs- und Vererbungslehre* 71, 156–230.

- Ganguly S, Barua D. 2020. High herkogamy but low reciprocity characterizes isoplethic populations of *Jasminum malabaricum*, a species with stigma-height dimorphism. *Plant Biol.* 22:899–909.
- Gilmartin PM. 2015. On the origins of observations of heterostyly in *Primula*. *New Phytol.* 208:39–51.
- Gutián J, Gutián P, Medrano M. 1998. Floral biology of the distylous Mediterranean shrub *Jasminum fruticans* (Oleaceae). *Nord. J. Bot.* 18:195–201.
- Gutiérrez-Valencia J, Fracassetti M, Berdan EL, Bunikis I, Soler L, Dainat J, Kutschera VE, Losvik A, Désamoré A, Hughes PW, *et al.*, 2022. Genomic analyses of the distyly supergene reveal convergent evolution at the molecular level. *bioRxiv:2022.05.27.493681*.
- Huu CN, Kappel C, Keller B, Sicard A, Takebayashi Y, Breuninger H, Nowak MD, Bäurle I, Himmelbach A, Burkart M, *et al.*, 2016. Presence versus absence of CYP734A50 underlies the style-length dimorphism in primroses. *Elife* 5:1–15.
- Huu CN, Keller B, Conti E, Kappel C, Lenhard M. 2020. Supergene evolution via stepwise duplications and neofunctionalization of a floral-organ identity gene. *Proc. Natl. Acad. Sci. U. S. A.* 117:23148–23157.
- Huu CN, Plaschil S, Himmelbach A, Kappel C, Lenhard M. 2022. Female self-incompatibility type in heterostylous *Primula* is determined by the brassinosteroid-inactivating cytochrome P450 CYP734A50. *Curr. Biol.* 32:671-676.e5.
- Kiew, R.; Baas, P. *Nyctanthes* is a member of Oleaceae. *Proc. Ind. Acad. Sci.* 1984, 93, 349–358.
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9:357–359.
- Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323.
- Li J, Cocker JM, Wright J, Webster MA, McMullan M, Dyer S, Swarbreck D, Caccamo M, Van Oosterhout C, Gilmartin PM. 2016. Genetic architecture and evolution of the *S*-locus supergene in *Primula vulgaris*. *Nat. Plants* 2:1–7.
- Lloyd DG, Webb CJ. 1992. The Evolution of Heterostyly. In: SCH Barrett, Editor. *Evolution and Function of Heterostyly*. Monographs on Theoretical and Applied Genetics. Vol. 15. Berlin (Germany): Springer.
- Matzke CM, Shore JS, Neff MM, McCubbin AG. 2020. The turnera style *S*-locus gene *TsBAHD* possesses brassinosteroid-inactivating activity when expressed in *Arabidopsis thaliana*. *Plants* 9:1–13.
- Matzke CM, Hamam HJ, Henning PM, Dougherty K, Shore JS, Neff MM, McCubbin AG. 2021. Pistil mating type and morphology are mediated by the brassinosteroid inactivating activity of the *S*-locus gene *BAHD* in heterostylous *Turnera* species. *Int. J. Mol. Sci.* 22.
- Olesen JM, Dupont YL, Ehlers BK, Valido A, Hansen DM. 2003. Heterostyly in the Canarian endemic *Jasminum odoratissimum* (Oleaceae). *Nord. J. Bot.* 23:537–539.
- Perrier de la Bâthie H. 1951. Notes biologiques sur les Oléacées de Madagascar et des Comores. *Mémoire de l'Institut Scientifique de Madagascar* 3: 175–186.
- Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139–140.
- Ryu, T.Y.; Yeam, D.Y.; Kim, Y.J.; Kim, S.J. Studies on heterostyly incompatibility of *Abeliophyllum distichum*. Seoul Natl. Univ. Coll. Agric. Bull. 1976, 1, 113–120.

- Sampson, D. R. 1971: Mating group ratios in distylic *Forsythia* (Oleaceae). *Canadian Journal of Genetics and Cytology* 13:368-71.
- Shore JS, Hamam HJ, Chafe PDJ, Labonne JDJ, Henning PM, McCubbin AG. 2019. The long and short of the *S*-locus in *Turnera* (Passifloraceae). *New Phytol.* 224:1316–1329.
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210–3212.
- Tange C (1998) *Silvianthus* (Carlemanniaceae) a genus and family new to Thailand. *Thai Forest Bull (Bot)* 26:59–65
- Taylor H. 1945. Cyto-taxonomy and phylogeny of the Oleaceae. *Brittonia* 5:337–367.
- Thompson JD, Dommée B. 1993. Sequential variation in the components of reproductive success in the distylous *Jasminum fruticans* (Oleaceae). *Oecologia* 94:480–487.
- Thompson JD, Dommée B. 2000. Morph-specific patterns of variation in stigma height in natural populations of distylous *Jasminum fruticans*. *New Phytol.* 148:303–314.
- Ushijima K, Nakano R, Bando M, Shigezane Y, Ikeda K, Namba Y, Kume S, Kitabata T, Mori H, Kubo Y. 2012. Isolation of the floral morph-related genes in heterostylous flax (*Linum grandiflorum*): The genetic polymorphism and the transcriptional and post-transcriptional regulations of the *S*-locus. *Plant J.* 69:317–331.
- Verdoorn IC. 1956. The Oleaceae of Southern Africa. *Bothalia* 6: 549–639.
- Xu S, Ding Y, Sun J, Zhang Z, Wu Z, Yang T, Shen F, Xue G. 2022. A high-quality genome assembly of *Jasminum sambac* provides insight into floral trait formation and Oleaceae genome evolution. *Mol. Ecol. Resour.* 22:724–739.
- Yasui Y, Mori M, Aii J, Abe T, Matsumoto D, Sato S, Hayashi Y, Ohnishi O, Ota T. 2012. *S*-locus EARLY FLOWERING 3 is exclusively present in the genomes of short-styled buckwheat plants that exhibit heteromorphic self-incompatibility. *PLoS One* 7:1–9.
- Young MD, Wakefield MJ, Smyth GK, Oshlack A. 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* 11:R14.
- Zhang C, Zhang T, Luebert F, Xiang Y, Huang C-H, Hu Y, Rees M, Frohlich MW, Qi J, Weigend M, *et al.*, 2020. Asterid Phylogenomics/Phylotranscriptomics Uncover Morphological Evolutionary Histories and Support Phylogenetic Placement for Numerous Whole-Genome Duplications. *Mol. Biol. Evol.* 37:3188–3210.

Appendix for Chapter 4

Supporting information includes the following items:

Figure S1. Number of transcripts according to different expression threshold.

Figure S2. Expression levels of differentially expressed transcripts in each samples.

Table S1. Sequencing statistics.

Table S2. Annotations information for differentially expressed transcripts.

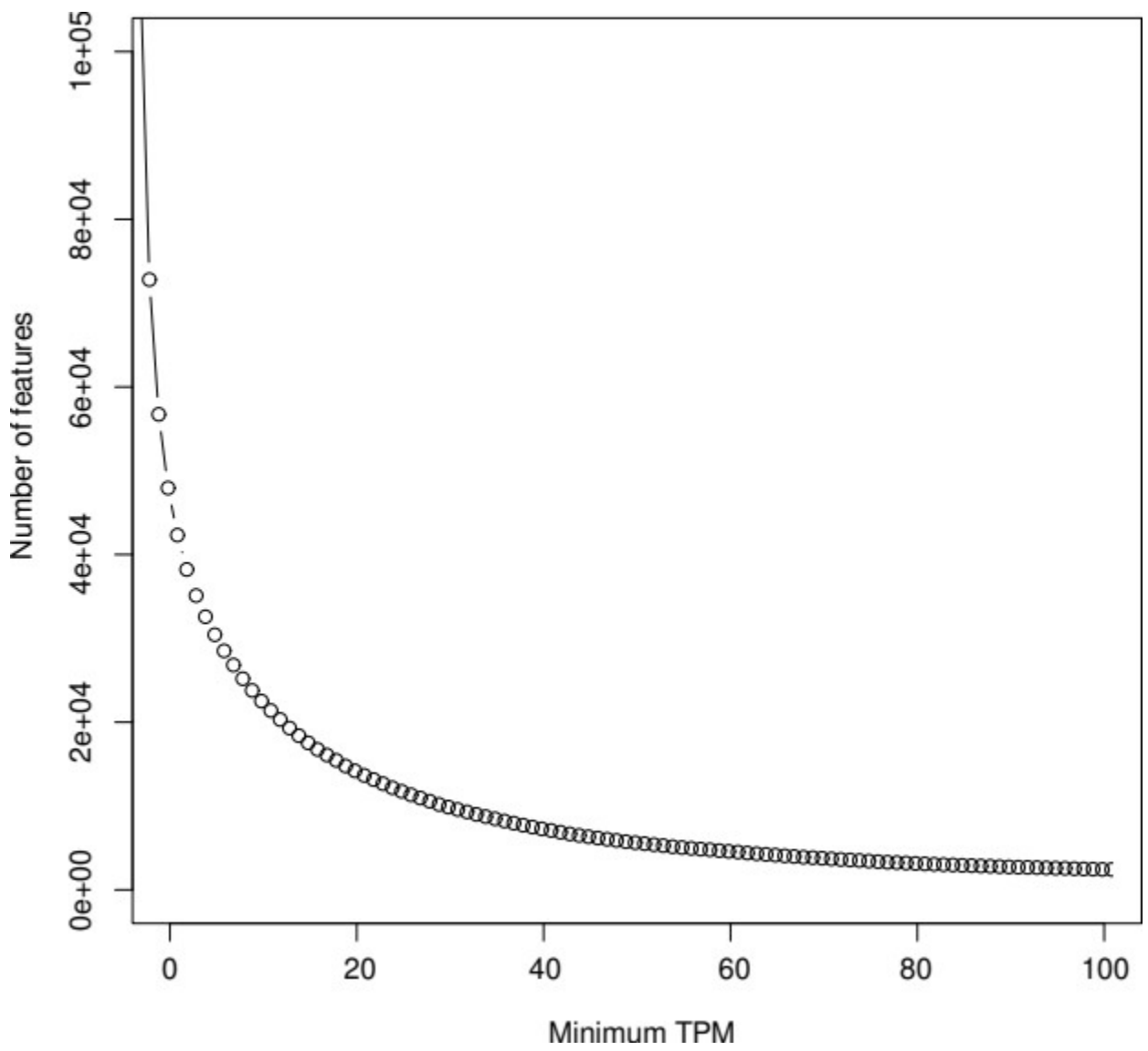


Figure S1. Number of transcripts according to different expression threshold.

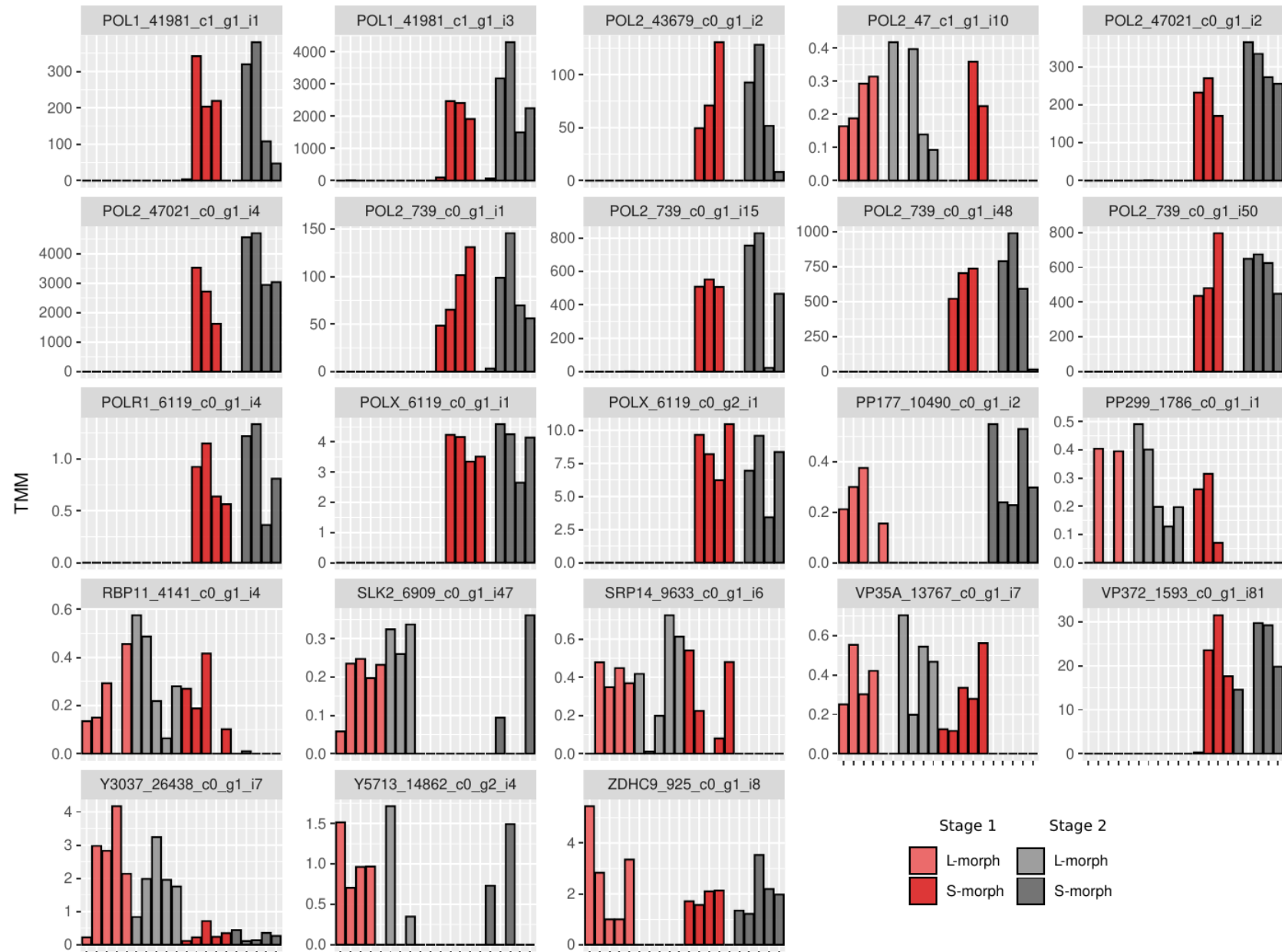
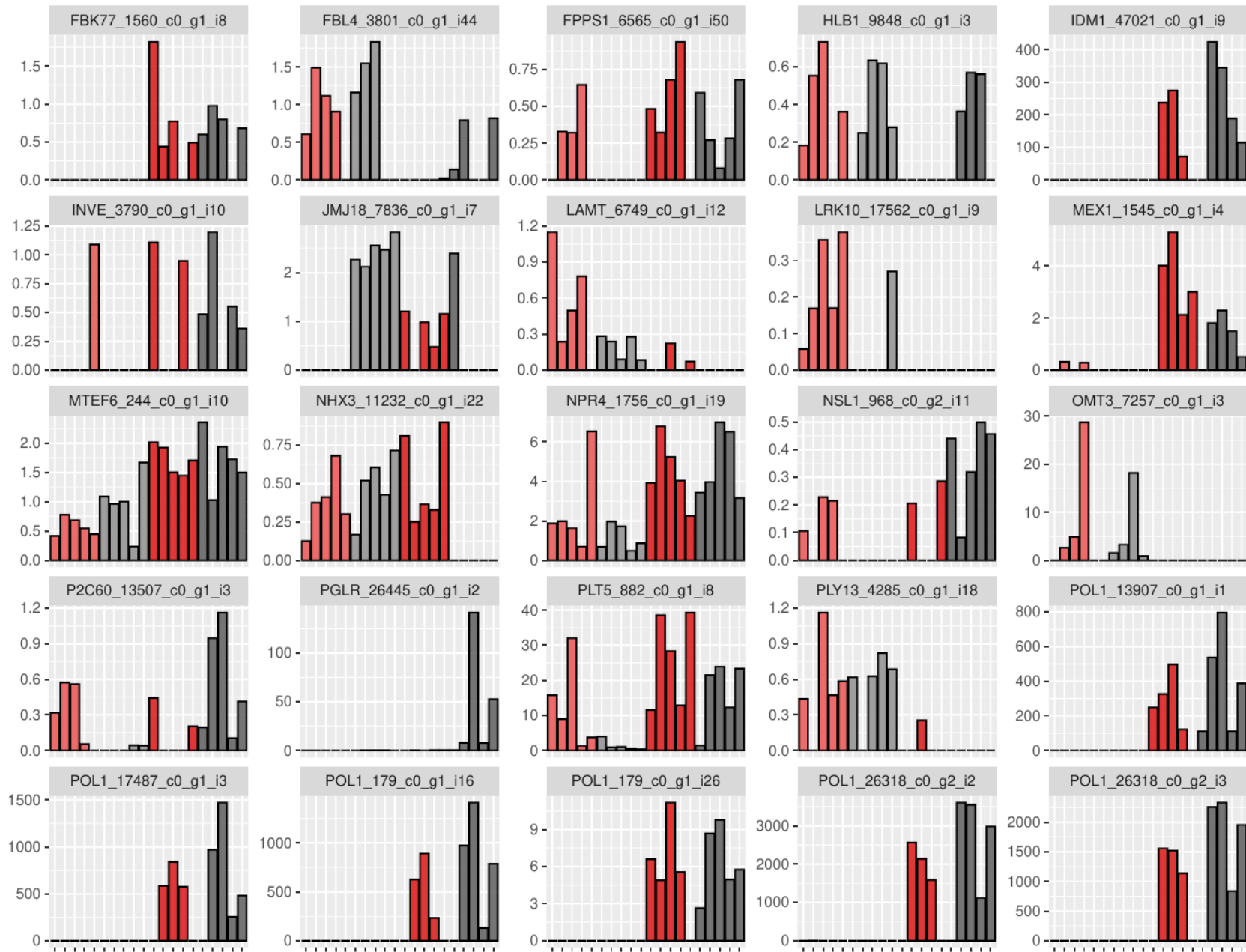


Figure S2. Expression levels of differentially expressed transcripts in each samples. Samples are colored according to developmental stage and darker shades denote S-morph individuals.



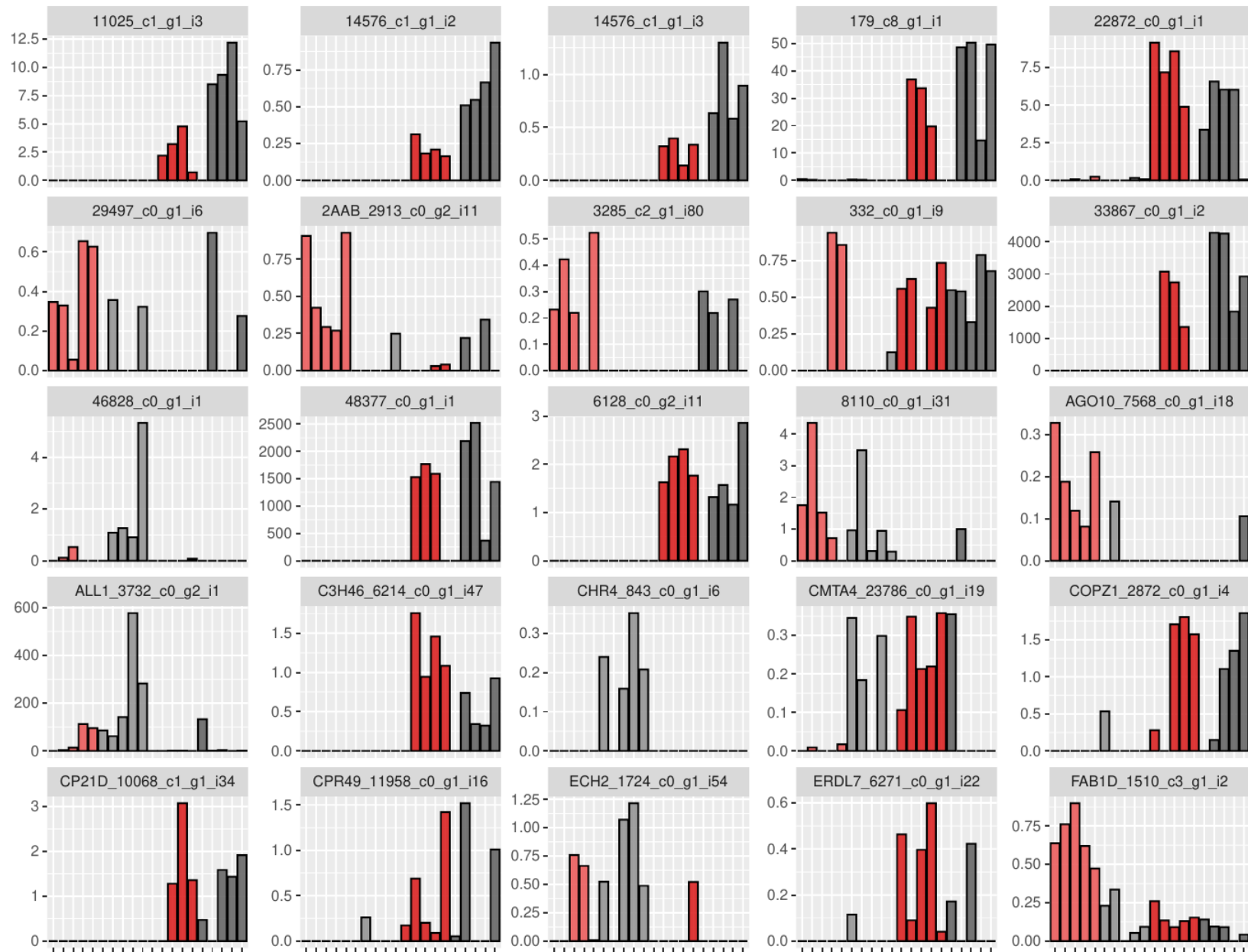


Table S1. Sequencing statistics.

Samples	Raw reads	Clean reads	Samples	Raw reads	Clean reads
L1-s1	40630168	38767596	S1-s1	36122886	34636886
L1-s2	32074356	30442554	S1-s2	42556292	41581176
L2-s1	44186372	42868186	S3-s1	40106330	39234010
L2-s2	33513128	32289076	S3-s2	41110408	40043220
L3-s1	39812570	38926896	S4-s1	41435572	40383596
L3-s2	41221704	39258636	S4-s2	40065948	39121278
L5-s1	48554478	47092296	S5-s1	43598944	42469720
L5-s2	40289054	38828608	S5-s2	38223332	37244744
L6-s1	34837080	33928108	S6-s1	38667680	37741838
L6-s2	35685230	34366410	S6-s2	46428456	45040018
Bigno-s1	40522314	39512096	Odora-s1	35800772	34774586
Bigno-s2	42687310	41403356	Odora-s2	39657478	38620110

Table S2. Annotations information for differentially expressed transcripts.

Transcripts	Annotation
TRINITY_DN179_c0_g1_i16	RNA1 polyprotein
TRINITY_DN739_c0_g1_i48	RNA2 polyprotein
TRINITY_DN739_c0_g1_i50	RNA2 polyprotein
TRINITY_DN739_c0_g1_i15	RNA2 polyprotein
TRINITY_DN739_c0_g1_i1	RNA2 polyprotein
TRINITY_DN33867_c0_g1_i2	-
TRINITY_DN13907_c0_g1_i1	RNA1 polyprotein
TRINITY_DN48377_c0_g1_i1	-
TRINITY_DN26318_c0_g2_i3	RNA1 polyprotein
TRINITY_DN41981_c1_g1_i3	RNA1 polyprotein
TRINITY_DN26318_c0_g2_i2	RNA1 polyprotein
TRINITY_DN179_c0_g1_i26	RNA1 polyprotein
TRINITY_DN47021_c0_g1_i9	RNA2 polyprotein
TRINITY_DN1593_c0_g1_i81	Vacuolar protein-sorting-associated protein 37 homolog 2
TRINITY_DN47021_c0_g1_i2	RNA2 polyprotein
TRINITY_DN47021_c0_g1_i4	RNA2 polyprotein
TRINITY_DN6119_c0_g1_i1	Retrovirus-related Pol polyprotein from transposon TNT 1-94
TRINITY_DN17487_c0_g1_i3	RNA1 polyprotein
TRINITY_DN6119_c0_g1_i4	Retrovirus-related Pol polyprotein from transposon RE1
TRINITY_DN6214_c0_g1_i47	Zinc finger CCCH domain-containing protein 46
TRINITY_DN26445_c0_g1_i2	Polygalacturonase
TRINITY_DN11025_c1_g1_i3	-
TRINITY_DN2872_c0_g1_i4	Coatomer subunit zeta-1
TRINITY_DN1560_c0_g1_i8	F-box/kelch-repeat protein At3g61590
TRINITY_DN41981_c1_g1_i1	RNA1 polyprotein
TRINITY_DN7836_c0_g1_i7	Lysine-specific demethylase JMJ18
TRINITY_DN43679_c0_g1_i2	RNA2 polyprotein
TRINITY_DN6119_c0_g2_i1	Retrovirus-related Pol polyprotein from transposon TNT 1-94
TRINITY_DN10068_c1_g1_i34	Peptidyl-prolyl cis-trans isomerase CYP21-4

TRINITY_DN14576_c1_g1_i3 -
 TRINITY_DN1545_c0_g1_i4 Maltose excess protein 1-like, chloroplastic
 TRINITY_DN6128_c0_g2_i11 -
 TRINITY_DN14576_c1_g1_i2 -
 TRINITY_DN179_c8_g1_i1 -
 TRINITY_DN3790_c0_g1_i10 Alkaline/neutral invertase E, chloroplastic
 TRINITY_DN6271_c0_g1_i22 Sugar transporter ERD6-like 7
 TRINITY_DN6565_c0_g1_i50 Farnesyl pyrophosphate synthase 1
 TRINITY_DN925_c0_g1_i8 Probable protein S-acyltransferase 7
 TRINITY_DN11958_c0_g1_i16 GDSL esterase/lipase CPRD49
 TRINITY_DN10490_c0_g1_i2 Pentatricopeptide repeat-containing protein At2g30780
 TRINITY_DN968_c0_g2_i11 MACPF domain-containing protein NSL1
 TRINITY_DN22872_c0_g1_i1 -
 TRINITY_DN23786_c0_g1_i19 Calmodulin-binding transcription activator 4
 TRINITY_DN13507_c0_g1_i3 Probable protein phosphatase 2C 60
 TRINITY_DN332_c0_g1_i9 -
 TRINITY_DN882_c0_g1_i8 Polyol transporter 5
 TRINITY_DN1756_c0_g1_i19 Rust resistance kinase Lr10
 TRINITY_DN1510_c3_g1_i2 Putative 1-phosphatidylinositol-3-phosphate 5-kinase FAB1D
 Probable inactive leucine-rich repeat receptor-like protein kinase
 TRINITY_DN26438_c0_g1_i7 At3g03770
 Serine/threonine-protein phosphatase 2A 65 kDa regulatory
 subunit A beta isoform
 TRINITY_DN2913_c0_g2_i11
 TRINITY_DN17562_c0_g1_i9 Ankyrin repeat-containing protein NPR4
 TRINITY_DN29497_c0_g1_i6 -
 TRINITY_DN1786_c0_g1_i1 Pentatricopeptide repeat-containing protein At4g01570
 TRINITY_DN4141_c0_g1_i4 Small RNA-binding protein 11, chloroplastic
 TRINITY_DN7568_c0_g1_i18 Protein argonaute 10
 TRINITY_DN6749_c0_g1_i12 Loganic acid O-methyltransferase
 TRINITY_DN6909_c0_g1_i47 Probable transcriptional regulator SLK2
 TRINITY_DN11232_c0_g1_i22 Sodium/hydrogen exchanger 3
 TRINITY_DN843_c0_g1_i6 Protein CHROMATIN REMODELING 4
 TRINITY_DN3801_c0_g1_i44 F-box/LRR-repeat protein 4
 TRINITY_DN9633_c0_g1_i6 Signal recognition particle 14 kDa protein
 TRINITY_DN4285_c0_g1_i18 Probable pectate lyase 13
 TRINITY_DN47_c1_g1_i10 Increased DNA methylation 1
 TRINITY_DN3285_c2_g1_i80 -
 TRINITY_DN13767_c0_g1_i7 Vacuolar protein sorting-associated protein 35A
 TRINITY_DN9848_c0_g1_i3 Protein HLB1
 TRINITY_DN3732_c0_g2_i1 Major pollen allergen Ole e 1
 TRINITY_DN46828_c0_g1_i1 -
 TRINITY_DN14862_c0_g2_i4 PI-PLC X domain-containing protein At5g67130
 TRINITY_DN1724_c0_g1_i54 Enoyl-CoA hydratase 2, peroxisomal
 TRINITY_DN7257_c0_g1_i3 Probable O-methyltransferase 3
 TRINITY_DN8110_c0_g1_i31 -

Chapter 5

Tracking the spatio-temporal spread of the cultivated olive with archaeogenomics

Raimondeau P., Wales N., Dulias K., Manzi S., Christin P.-A., Barazani O., Bouby L., Dag
A., Figueial I., Galili E., Kaniewski D., Newton C., Otto T., Pagnoux C., Rovira N.,
Terral J.-F., Tillier M., Wagner S., Orlando L., Besnard G.

Abstract

Olive is a Mediterranean tree of major cultural and agricultural importance. Olive cultivation is thought to have started approximately 6000 years ago in the Levant but many uncertainties remain about its precise domestication process. Archaeogenetics offers great potential to bring new insights on the crop origins, diffusion, and secondary diversification. Using aDNA extracted from olive stones originating from diverse archaeological sites around the Mediterranean Basin (from the Neolithic to the Roman period), we charted the dispersal of DNA lineages through space and time. Our methodology relies on the comparison of archaeological samples to modern references. This comparison, done at both nuclear and organellar level allowed us to assign most samples to a specific olive modern lineage. The dominance of a unique organelle haplotype in ancient samples supports a scenario with a main olive domestication event in the eastern Mediterranean area. Our results also revealed that the most widespread extant cultivated olive lineage was already cultivated in France during the Antiquity. This work constitutes the first demonstration of the use of archaeogenomics to uncover olive domestication history.

Key-words: ancient DNA, *Olea europaea*, domestication, organellar genomes

Introduction

The olive (*Olea europaea* L. subsp. *europaea*) is an iconic tree of the Mediterranean cultural heritage, being a landmark of the Mediterranean vegetation but also providing essential ingredients of the Mediterranean food. The multiple uses of olive explain its expansion with the spread of some of the most ancient civilizations (Kaniewski *et al.*, 2012). Its importance has motivated a great deal of research to decipher the domestication history of this species from paleobotanical, archaeological and molecular data (see Besnard *et al.*, 2018 for review). While the precise timing and locations are still discussed, it is widely accepted that olive domestication started approximately 6000 years ago in the eastern Mediterranean basin (Lipshitz *et al.*, 1991). Among the six olives subspecies, the main wild progenitor of the cultivated olive is the wild Mediterranean olive, also called oleaster (Angiolillo *et al.*, 1999; Besnard *et al.*, 2007). Two main oleaster gene pools have been identified in the western and eastern Mediterranean Basin (Besnard *et al.*, 2001, 2013b; Breton *et al.*, 2006; Belaj *et al.*, 2007, 2011; Díez *et al.*, 2015). Nuclear markers demonstrated a stronger affiliation of the cultivated germplasm to wild individual from the East, with a significant contribution of western oleasters (Besnard *et al.*, 2001; Gros-Balthazard *et al.*, 2019; Julca *et al.*, 2020). It is unclear if this affiliation results from one or multiple local domestication events. In the first scenario, a second gene pool was independently domesticated in the central Mediterranean basin (Díez *et al.*, 2015). In

the second scenario, cultivars from the East were spread to westernmost regions of the Mediterranean basin, where they were introgressed with local wild populations allowing secondary diversification (Besnard *et al.*, 2018). The diffuse character of olive domestication process as well as the overlapping distribution of oleasters and cultivated trees, often difficult to distinguish, have hampered setting the matter.

The recovery and analysis of DNA from archaeological plant remains has provided valuable insights about domestication of annual crops such as maize (Da Fonseca *et al.*, 2015; Ramos-Madrugal *et al.*, 2016; Pérez-Zamorano *et al.*, 2017), barley (Mascher *et al.*, 2016) or sunflower (Wales *et al.*, 2019) and, more recently, perennial plants such as grape (Ramos-Madrugal *et al.*, 2019) or date palms (Gros-Balthazard *et al.*, 2021; Pérez-Escobar *et al.*, 2021). The preservation of DNA in olive pits was demonstrated by Elbaum *et al.*, (2006), and the few sequences produced in this study remain to date the only archaeogenetics information for olive. Ancient DNA (aDNA) thus offer great potential to bring new insights on the cultivated olive origins, diffusion, and secondary diversification by providing anchor points for phylogeographic studies for instance.

Mitochondrial and chloroplast DNA polymorphisms are markers of choice for phylogeographic reconstructions of the olive tree (Besnard *et al.*, 2001; Hong-Wa and Besnard, 2013a; Van de Paer *et al.*, 2018). Indeed, organellar genomes are maternally inherited in *O. europaea* (Besnard *et al.*, 2000), thus only spread by seeds, at shorter distance than nuclear genes. These haploid organellar genomes are also more prone to stochastic events because their effective population size is half that of diploid genomes, allowing a more accurate detection of evolutionary events such as a long persistence of relict populations in refuge zones or post-glacial recolonization (Petit *et al.*, 2002). Moreover, their presence as multiple copies increases chance of recovery and availability for a lower sequencing effort. In the wild and cultivated Mediterranean olive, three cytoplasmic lineages have been described: E1 originates from the eastern Mediterranean region (where it is highly diversified) and has been spread with the cultivated olive during historical times, while E2 and E3 are specific to the western area (from the Peloponnese and Cyrenaica to the Iberian Peninsula and Morocco). The three lineages have diverged since at least the Early Pliocene and their diversification is relatively recent, since the Late Pleistocene (Besnard *et al.*, 2013b). As a consequence, each plastid lineage coalesces rapidly, allowing unambiguous identification.

Using aDNA extracted from olive stones and wood originating from diverse archaeological sites mostly around the Mediterranean Basin (from the Neolithic to the Roman period), we aimed to chart the dispersal of DNA lineages through space and time. We first focused on organellar DNA and gathered an extensive panel of modern divergent olive haplotypes. This collection was used to

constitute a reference set of diagnostic single nucleotide polymorphism (SNP) markers for distinct olive lineages. To investigate the ancestry of ancient samples, we also evaluated the relationship between ancient and modern accessions using nuclear data. This strategy allowed us to pinpoint most archaeological samples to a modern lineage. This information revealed that the major extant cultivated olive lineage was already cultivated in southwestern Europe 2000 years ago. Its abundance among ancient samples points toward a main domestication event of the olive in the eastern Mediterranean Basin.

Material and methods

Sampling and radiocarbon dating

We collected 25 olive stones from nine archaeological sites around the Mediterranean Basin (Figure 1A), as well as wood from a lancet from Saruq in Southeast Arabia (Table S1). According to archaeological evidence, samples date from the Neolithic to the Roman period (ca. 6500-500 BP) and come from humid contexts, except for the wood sample. We checked date estimates for four samples by radiocarbon dating carried out at the University of Arizona AMS facility. DNA was extracted from olive endocarps and wood in a dedicated aDNA facility at the University of York using a method developed for archaeobotanical remains (Wales et al., 2019).

DNA extraction and sequencing

The recovered DNA was converted to double-stranded DNA libraries following the protocols of Wales et al. (2019). Libraries were pooled and sequenced on one lane of HiSeq2000, including two extraction controls. We later sequenced more deeply three of the libraries (MTF1, MTF7 and MTC79) on a lane of Novaseq.

DNA content and quality

Endogenous DNA content of samples as well as DNA damage levels were then evaluated using Paleomix (Schubert *et al.*, 2014) and MapDamage (Jónsson *et al.*, 2013). In short, for each sample reads were trimmed to remove low quality bases and mapped onto the wild olive nuclear reference genome (Unver *et al.*, 2017). Mapped reads were used as an approximation of the endogenous DNA content and level of clonality. MapDamage then use substitutions information to perform Bayesian estimation of damage parameters and model post-mortem DNA damages such as misincorporation patterns.

Generating a modern organellar SNP panel

We gathered plastome and mitogenome sequences for 25 and 21 modern olive accessions, respectively (Table S2). They represent distinct haplotypes of Mediterranean olive lineages (*O. e.* subsp. *europaea*; 9 E1, 6 E2, 4 E3) and seven lineages of other subspecies (C1.1, C1.5 and C2 for *O. e.* subsp. *cuspidata*, L1.6 for *laperrinei* and M3.3 for *guanchica*). Fourteen of them were newly generated in this study according to the protocols described in (Van de Paer *et al.*, 2018). We used haplotype E1.2 (Besnard *et al.*, 2013b) as the primary reference for organellar genomes study.

Each modern sequences for mitochondria and plastid was hashed into random reads without errors with BBTools (Bushnell B. - sourceforge.net/projects/bbmap/) with a target coverage of 20. Reads were then mapped with Bowtie2 (Langmead and Salzberg, 2012) on the E1.2 plastome (after removal of one inverted repeat) and mitogenome simultaneously. Using reads instead of directly aligning genomes avoids the creation of gaps in the reference due to indels. By conserving the base numbering of the reference genome, it was then possible to directly compare bases called at a given position in ancient samples to this reference SNP sets. In order to avoid calling SNP in *mtpt* (plastid sequences transferred into the mitogenome (Van de Paer *et al.*, 2018) and repetitive regions, which may hinder the lineage assignment by creating conflictual results, multi-mappers were excluded (by keeping only reads with a mapping quality greater than 20 with samtools v1.9 (Danecek *et al.*, 2021). Variant calling was performed with vcftools v0.1.16 (Danecek *et al.*, 2011), keeping only biallelic SNPs with a Phred score quality greater than 20 and excluding sites within 3 bp of indels. We finally filtered out sites with a depth greater than twice the mean depth to exclude errors due to mapping in paralogous/repetitive regions. The remaining variants constituted our reference organellar SNP panel.

Identifying SNPs in archaeological samples

Adapters were trimmed with cutadapt (Martin, 2011) and we used the bam_pipeline implemented in Paleomix to map reads from each sample on olive references (nuclear, plastid and mitochondrial). As part of the pipeline, reads were mapped with BWA without seeding (Li and Durbin, 2009), with realigning around indels performed with GATK (McKenna *et al.*, 2010). MapDamage was used to model DNA damages and rescale quality score accordingly. For the three samples re-sequenced to a greater depth with paired-end reads, overlapping pairs were merged prior to alignment to improve alignment precision.

For organellar genomes, we called SNP the same way we did for modern samples but only kept those also presents in our reference SNP dataset. Restricting the SNP calling to variable

positions detected in published genomes minimizes the risk of false positives due to sequencing errors and DNA damage, especially as we are working with low-depth sequences.

For the nuclear genome, we used Bowtie2 (default parameters) to map reads from modern WGS samples for 14 oleasters (8 and 6 from eastern and western gene pools, respectively; Table S2) on the olive reference genome (Unver *et al.*, 2017). We also used STAR2 (default parameters; (Dobin *et al.*, 2013) to map oleaster RNA-seq samples (9 eastern, 4 western; Table S2) and 4 *O. e. cuspidata* samples to use as outgroup. Modern and ancient samples alignment files were processed together in ANGSD (Korneliussen *et al.*, 2014) and we directly used genotype-likelihoods for downstream analyses. We estimated minor and major allele frequencies (option -doMajorMinor 1) using the GATK GL model (option -GL 2), a base quality of 30 (option -minQ 20), and a minor allele frequency of 0.05. To reduce the amount of missing data stemming from the different nature of the modern genomic data (i.e. whole-genome sequencing or RNA-seq), we only used coding sequences by restricting the analyses to coding position according to the reference genome annotation file.

Assignment of archaeological samples to modern cytoplasmic lineages

For each sample, we then computed a similarity score to quantify the affiliation strength of each ancient sample to a given haplotype. To do so, we coded reference, alternative and missing alleles as -1, 1 and 0, respectively. Sites were weighted according to their specificity among the accessions in our reference panel, by dividing the site code by the number of accessions harbouring this same variant. For a given sample, the similarity score for a given haplotype was thus the mean of all the weighted-site values over informative sites in this haplotype. Samples scores were then scaled to the maximal reachable score for each haplotype (if all alleles were typed and alternative). This maximum score indicates the level of information available to pinpoint a lineage. The smaller it is, the less informative the haplotype score will be (no variable sites, or solely shared with other haplotypes). Conversely, a maximum score close to 1 denotes a haplotype with haplotype-specific SNPs. We proceeded similarly for both organellar genomes and then used the distance between the two pairwise similarity matrices and performed a Mantel test (1000 permutations) with the R package *vegan* (Oksanen *et al.*, 2022) to evaluate their congruence. The difference between maximum scores for both organelles indicates their relative power to identify a given lineage.

Phylogenetics and population genetics analyses

For organellar genomes, we generate consensus sequences for each modern and ancient sample from the aligned reads using the consensus tool in ANGSD. That way all plastid and mitochondrial sequences were already aligned and we did not introduce a bias between ancient and modern

samples by using a reference only for the ancient samples. We removed individuals with more than 60% missing data and used IQtree2 (Minh *et al.*, 2020) to build phylogenetic trees from mitochondrial and plastid sequences separately with a GTR model and 1000 ultrafast bootstrap pseudoreplicates.

On the nuclear dataset, we carried out a principal component analysis (PCA) with ANGSD including the modern and the four archaeological samples with a coverage greater than $0.1\times$ using the single read sampling strategy (-doIBS 1) and excluding all transitions to account for potential remaining deaminated bases in the aDNA samples. To investigate potential admixture in ancient samples, we also performed clustering analysis with the admixutre module of ANGSD, with K ranging from 2 to 5. The explanatory power of the increasing K values was assessed using the ΔK criterion (Evanno *et al.*, 2005), implemented with the CLUMPAK tool (Kopelman *et al.*, 2015). To dig into potential introgression in one ancient sample (MTC79), we used the ABBA-BABA test as implemented in ANGSD. This test is based on the counts of the number of bi-allelic sites that have a topology different from the four-species tree. In the event of gene flow, a bias in the proportion of ABBA and BABA sites is expected. We considered Z scores higher than 3 or smaller than -3 to reject the null hypothesis. The individuals considered for this analysis were the MTC79 sample, MZ4 and OS4 representing the western and eastern gene pool respectively as well as OB4 (*O. e. cuspidata*) as an outgroup.

Results

Olives stones up to 6000 years old treasure endogenous DNA

Out of the 28 archeological samples, the six extracts with the lowest endogenous DNA content (0.05%) did not produce enough read to fit DNA damage models (Table S3). The three samples from Tweini provided sequences of suspiciously great quantity and quality (Figure 1C). We thus decided to use radiocarbon dating on one of these samples which confirmed a modern origin. Excluding these nine problematic samples, endogenous content varied between 0.1 and 17.3% and sequence data displayed typical aDNA damage patterns (Supplementary Dataset 1); i.e. enrichment in transitions, particularly at fragment extremities due to deamination (Hansen *et al.*, 2001). As previously reported for ancient plant samples (Staats *et al.*, 2011), plastid DNA generally harbors higher damage levels than mitochondrial and nuclear levels, although some confidence intervals overlap. This could be explained by a greater detection sensibility on plastid DNA due to the greater read depth. Older samples generally exhibited higher damage levels and lower DNA contents (Figure 1; Supplementary Dataset 1). The sequencing depth was the best for plastid DNA with 13 samples above $1\times$ (up to $18\times$). Eight samples were sequenced to more than $1\times$ on the mitogenome

(up to 4×). Finally, four samples had a depth of coverage greater than 0.1× on the nuclear genome (Table S3).

Organellar genomes variants circumscribe maternal haplotypes

We first focused on organellar DNA and gathered an extensive panel of divergent olive haplotypes to constitute a reference set of diagnostic SNP markers for non-Mediterranean subspecies (*cuspidata/laperrinei/guanchica*), major lineages in *Olea europaea* subsp. *europaea* (E1/E2/E3) as well as for distinct haplotypes within these lineages. In comparison with E1.2 references, we counted between 1 and 89 SNPs in plastome and 5 to 304 in mitogenome. Many of these variants are common between several samples. We identified 278 and 751 distinct SNPs in plastid and mitochondria, respectively. The elevated number in mitochondria only reflect the difference in size between the two genomes, as it represents 10.5 variants for 10,000 bp whereas the plastid rates was twofold with 21.4 variants for 10,000 bp. This proportion is congruent with the differential rate of evolution between the two genomes (Drouin *et al.*, 2008). Mitogenome and plastome also differed in their level of informativeness with the plastome showing 51% of haplotype-specific variants and the mitogenome only 34% (Figure 2). In both genomes, the greatest number of unique variants was detected in subspecies *cuspidata*, followed by lineages E3 and E2 of *europaea*, *guanchica* and finally *laperrinei*. This is consistent with the expectation based on the phylogeny of the olive complex (Van de Paer *et al.*, 2018). This order is not conserved when considering mitochondria and non-unique variants, showing homoplasmy can mislead identification at this level.

Archaeological samples can be assigned to modern organellar diversity

We called up to 386 and 750 SNPs in plastid and mitochondria genomes, respectively (Table S4). However, most of them were not in the reference panel (maximum 2 in plastome, 24 in mitogenome). We retrieved SNPs in all samples, however, seven of them did not provide any SNP common with our panel, hindering their assignment to a subspecies or a cytoplasmic lineage (all samples from Saruq and Samos, plus one sample from Mount Carmel). We were only able to exclude affiliations to a few haplotypes for two additional samples from Mount Carmel (Figure 3). Combining information from both organelles, we were able to deem all the remaining samples (19/28) as most likely belonging to lineage E1. We do note that four of these assignments were based on low evidence level (TW4, TW2, PAP5 and CAS3). The proportion of sites from the panel that were genotyped in these 19 samples ranges from 4 to 95% (in TW4 and MTC79, respectively) with a median of 68% for plastome and 28% of typed sites for mitochondria (Table S4). Plastid DNA overall provided more information than mitochondrial DNA and three samples that could not

be assigned to a lineage based on the latter were successfully assigned based on the plastome information. The distances between the two matrices, showed the signal from plastome and mitogenome were globally congruent. The two matrices are significantly correlated (Mantel test p -value < 0.001; $R^2 = 0.85$). There was only conflicting affiliation between mitochondria and plastid regarding E1.4 haplotype, and mitochondria has a way greater affiliation power in this case (Figure S1). Indeed, for E1.4 affiliation in plastome, the power is nearly null as there is only one informative site and it is shared with 15 haplotypes. On the other hand, there are seven E1.4 specific variants in our mitochondrial panel. Similarly, E1.3, E1.5, E1.10 and E1.13 show the same plastid profile than E1.4 with only one non-unique variants distinguishing them from E1.2 in our panel. Distinguishing between these six haplotypes based on plastid information is thus not possible. In addition, in the mitochondrial variants panel, E1.3, E1.5 and E1.10 do not harbor any variable sites compared to E1.2.

We were able to refine our affiliation to specific E1 sub-lineages for 10 of the 19 samples. These affiliations are not strict but only indicates a greater proximity with one of the reference haplotypes. We identified eight samples (TW1, SAU5, SAU4, SAU3, PAP4, LCH, JV141, JV131) as belonging to haplotype E1.1 with great confidence. MTF1 and MTF6 show proximity to E1.1. For MTF7 and MTF2 a different profile is suggested, as numerous genotypes are not congruent with E1.1 in both plastome and mitogenome. Some E1.1-specific sites are also found in the mitogenome, making the assignment of these two samples tricky. Information from both genome attest that MTC79 is not E1.1 but more likely E1.2 (or E1.3/5/10 as they can't be discriminate in our panel). We also look at two length-polymorphisms that allow distinguishing the three Mediterranean lineages: a 243-bp deletion in *ycf1* specific to E3, and an 8-bp deletion (ATTAGATA repeated once or twice) in *trnS-trnG* specific to E2 (Besnard *et al.*, 2011). Reads were not found at this position in *ycf1* for samples from Saruq and Samos, MTC78/14/15, CAS3, TW2 and TW4. For the remaining 16 samples, the presence of read(s) allow their assignment to E1 or E2. For *trnS-trnG*, we looked at read(s) spanning the whole indel due to its repetitive nature. If a single mapped read contained twice the repeat, a E2 haplotype could be excluded. It was the case for only five samples (that also harbor *ycf1* reads): MTF7 and 1, SAU3 and 5, and TW1. Combining these new pieces of information with the results from SNP calling, we obtained congruent combinations pointing to an E1 haplotype for all samples.

Eight samples provided enough data to be included in plastid DNA phylogeny to investigate their relation to other haplotypes (Figure 4). Confirming our SNP based affiliation, close relationship with E1.1 was observed for all but two samples; MTC79 (Mount Carmel, Israel) and MTF7 (Montferrier, France). Nevertheless, other samples from Montferrier (MTF6 and MTF1)

were more closely related to E1.1 than to MTF7. We note that branches in ancient samples are notably long compared to modern samples, probably enriched in sequencing errors. Some placement may then be artefactual, which seems even more likely in the mitochondrial phylogeny where all ancient samples group together (Figure S2).

Nuclear ancestry of archaeological samples

Unlike the plastome, which only traces one genetic lineage (maternal), the relative proportion of eastern and western ancestry can be estimated from the nuclear genome, eventually allowing to arbitrate between different domestication scenarios. We only analyzed the four samples with the greatest depth ($>0.1 \times$ MTC79, TW1, MTF7, MTF1). In the PCA, they all clustered with the eastern modern oleasters, yet MTC79 was a bit distant from other samples (Figure 5A). The modern samples the most distant from other members of their genepools on the two first axes of the PCA (Oeiras and Kitries) are known to be admixed based on nuclear microsatellites (Besnard *et al.*, 2013a). Clustering results were consistent with this observation and show some signal of introgression from Western olive in MTC79 (Figure 5B). Nonetheless, the ABBA-BABA test results were not significant. This might be explained by a lack of power due to the low number of sites available.

Discussion

Ancient DNA from olive stones is exploitable

While recovering DNA from ancient source is becoming more common, it is still not routine, and this project represents the first of its potential use to understand the domestication of olive. We demonstrated that retrieving informative DNA from waterlogged olive stones is feasible. In contrast, our wood samples did not contain a high enough DNA content to be informative (Table S1, S2). This is probably due to conservation condition as DNA from wood samples have successfully been recovered in other species (Dumolin-Lapègue *et al.*, 1999; Liepelt *et al.*, 2006; Wagner *et al.*, 2018). We chose to only use variants that were previously identified in modern samples. This strategy greatly reduced the amount of information available and might exclude genuine variation present in ancient samples reflecting unknown diversity. However, separating the variant discovery process from the variant genotyping process allow relaxing stringency on the latter which is of particular interest in the case of aDNA where information is scarce and prone to errors. One read can then be a sufficient evidence for the presence of an alternative allele if it has already been characterized in an independent dataset. This Bayesian approach is common in genotyping process for low-coverage data (Korneliussen *et al.*, 2014). As illustrated by the

inclusion of archaeological samples in phylogenetics trees (Figure 4, Figure S2), the amount of unidentifiable variants present in aDNA is more likely to dilute useful information than to reveal new one.

Organellar genomes are a privileged source of information

Using organellar information, we assigned 19 of our archaeological samples to a unique Mediterranean olive lineage: E1 (Figure 3). Congruent results were obtained between both genomes but we note that in the case of olive, mitochondrial data are generally less informative and prone to homoplasy (Figure 2, Figure S1). When haplotype-level affiliation was possible, all but one of the archaeological samples were assigned to the haplotype E1.1. This haplotype accounts for 80% of present-day cultivars. This result demonstrates that it was already introduced during the Antiquity in the Greek city of Massalia, in southern France (Figure 1). Beyond the idea that extant olive cultivars pedigrees take root in olive cultivated 2000 years ago, the dominance of E1.1 in archeological samples of various origins point towards a scenario with a major domestication event in the eastern Mediterranean. Indeed, in the event of multiple local domestication, we would expect to find multiple plastid lineages. The oldest sample we were able to assign to an extant haplotype MTC79 was however ascribed to E1.2 (or E1.3/E1.5/E1.10). Nowadays, E1.2 is the second most frequent cultivated haplotype (7%; Besnard *et al.*, 2013b), particularly common in Egypt and the Levant. Two other samples were not assigned to E1.1 but show mixed profiles, between E1.1 and E1.2 that would require further investigations.

Future of archaeogenetics in olive

While organellar genomes are a great resource to track the spread of lineages during domestication, the most interesting questions remaining about olive domestication require the use of nuclear information. To further investigate how archaeological samples relate to extant ones, we attempted to determine nuclear genome ancestry of the ancient samples with the best coverage. One of these samples, MTC79, originated from a submerged site off the Carmel coast in Israel that was estimated to date from 6500 years ago (Galili *et al.*, 2021). It is unclear if it predates domestication or represents early dates cultivated olives. Its genetic characterization is thus of great interest. The genetic information we gathered points toward a filiation to the eastern genepool as expected, but we detected some conflicting patterns suggesting there might be admixture with western olives (Figure 5). It might reflect some exchanges between East and West Mediterranean during the Neolithic, when environmental conditions were more humid in North Africa and the Levant (reviewed in Médail and Quézel, 2018) and thus favourable for the spread of the olive (e.g. Besnard

et al., 2013a) and other cultivated trees such as date palms (Gros-Balthazard *et al.*, 2021). More investigation would be needed to determine the reliability of this observation.

A promising way to investigate introgression profiles in ancient samples would be to identify and quantify distinct repeat families between western and eastern gene pools. It has been shown that repeat content can vary between distinct olives lineages (Contento *et al.*, 2002; Mascagni *et al.*, 2022). The higher abundance of nuclear repetitive DNA could allow the characterization of samples with low coverage, as the ones we generated. For instance, RepeatExplorer (Novák *et al.*, 2020) is a tool designed to identify and quantify repeats from unassembled NGS reads with the rationale that if one sequences randomly a genome the repetitive regions will represent a higher proportion of reads than the single-copy regions. This approach has however never been tested on ancient DNA samples and might be hindered by a poor recovery of methylated DNA sequences such as pericentromeric satellites (Brändle *et al.*, 2022). However, transposable elements contents were successfully used to investigate cotton domestication (Palmer *et al.*, 2012), suggesting it may be a viable approach in olive, providing a clear distinction can be made between transposable element families of the western and eastern gene pools.

Other particularly interesting questions could be addressed with a more thorough sequencing effort. One can ask which traits were of primary interest during domestication and when. For instance, study of two archaeological sites on the Carmel coast of Israel (including the one from which our samples originated) concluded in different usages of olive for table olive or oil production. Without access to genetic information, the sequence of character selection is only accessible from morphological features such as fruit shape (Terral *et al.*, 2004). This type of study can uncover surprising selection targets, as illustrated by a recent study on watermelon domestication that demonstrates from archaeological samples, selection for seeds not flesh (Pérez-Escobar *et al.*, 2022). Ancient nuclear genomes for chosen ancient olives would allow to track sequence changes in a wide range of genes linked to phenotype over time. The integration of archaeogenetics data opens an exciting chapter in the study of olive domestication.

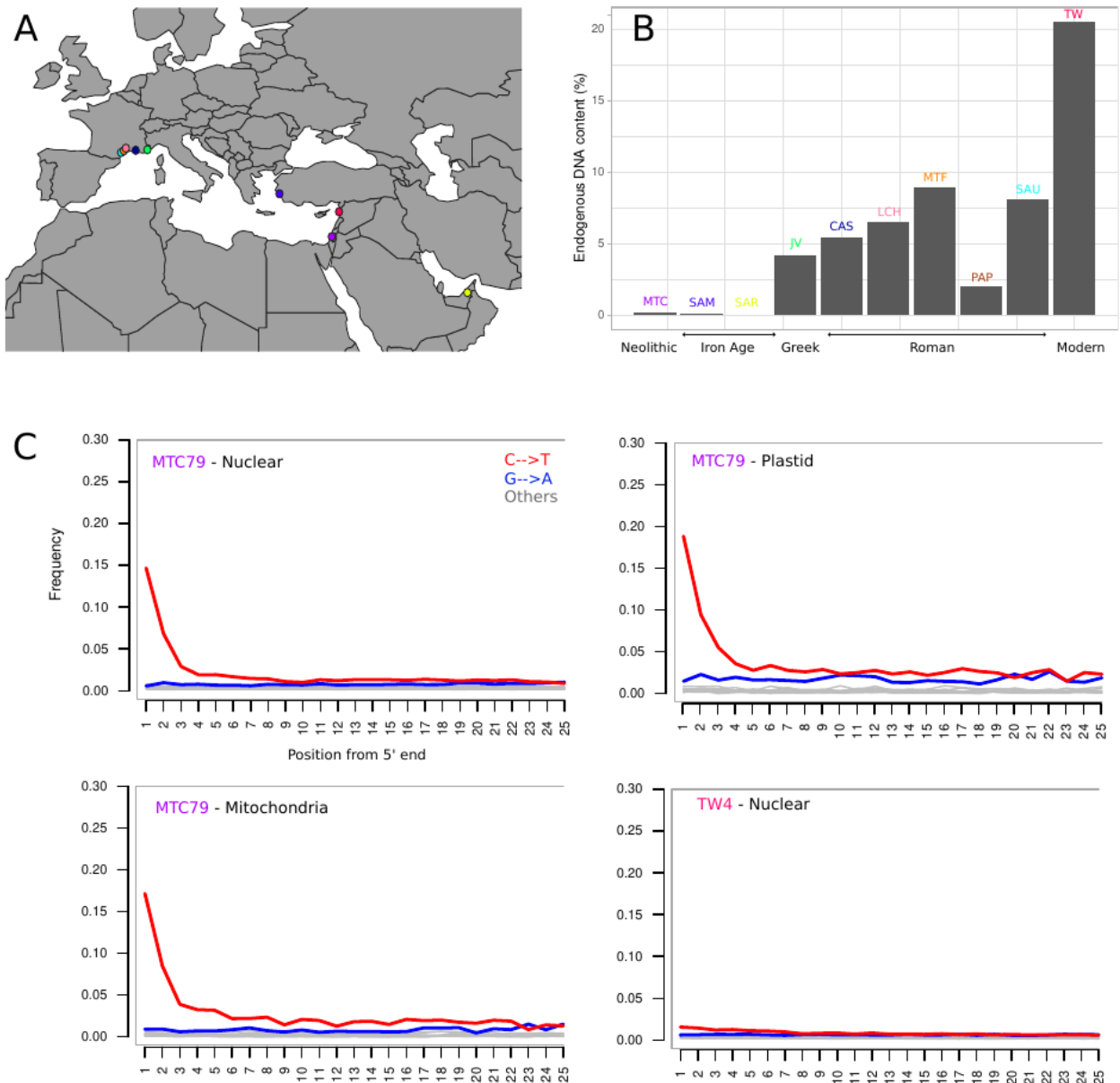


Figure 1. Ancient DNA samples characteristics. A) Locations of archaeological sites where the samples were collected. Each site is represented with a distinct colour, used in other panel. B) Median endogenous DNA content of samples from each site, organized by age. C) Patterns of DNA substitutions at 5' end for MTC79 (from Neolithic) plastid, mitochondrial and nuclear DNA and for TW4 nuclear DNA (modern). C to T substitution are shown in red, G to A in blue, grey lines represent other types substitutions. For complete profiles for all samples, see Supplementary Dataset 1.

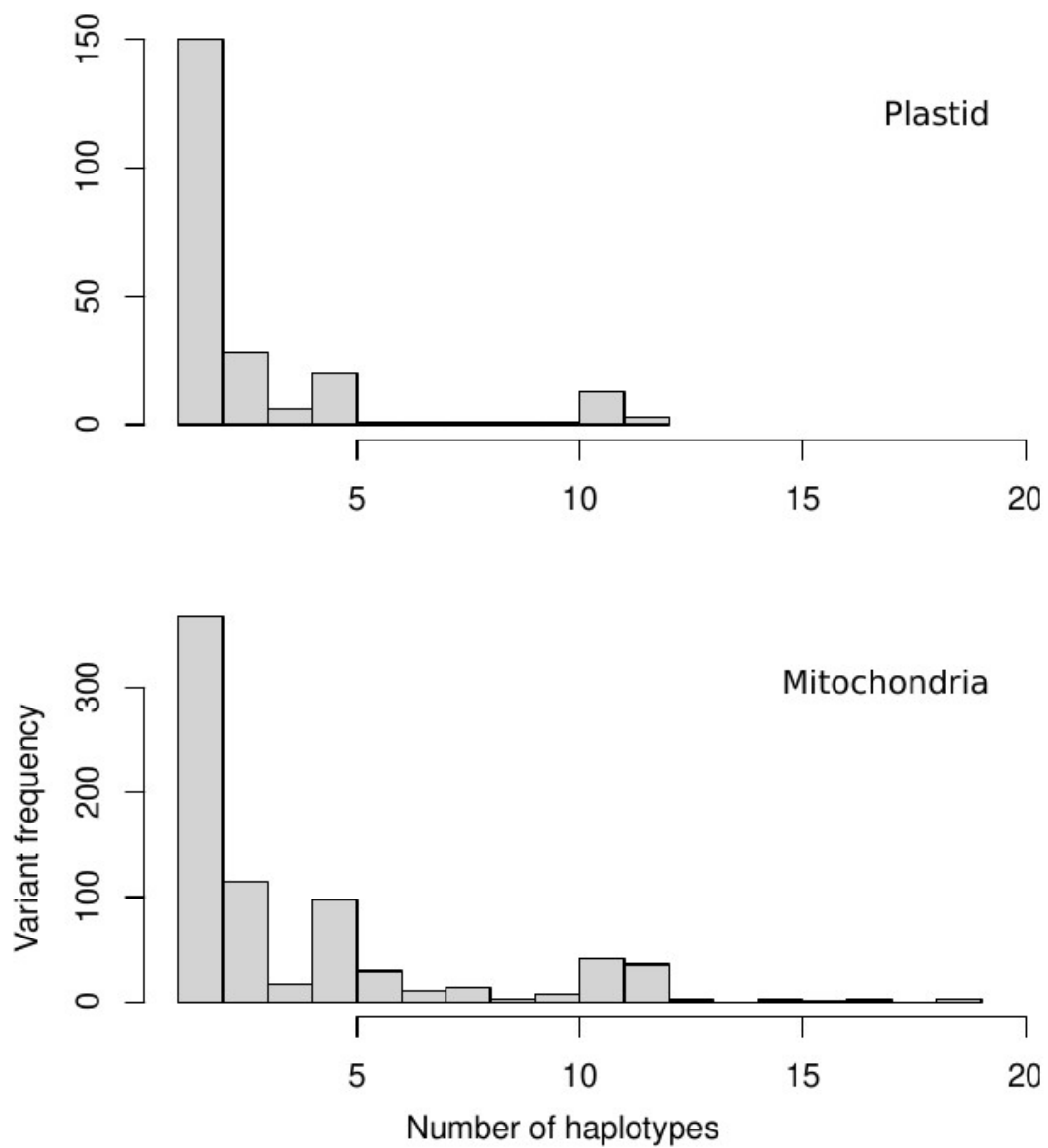


Figure 2. Distribution of modern plastid and mitochondrial variants among reference haplotypes.

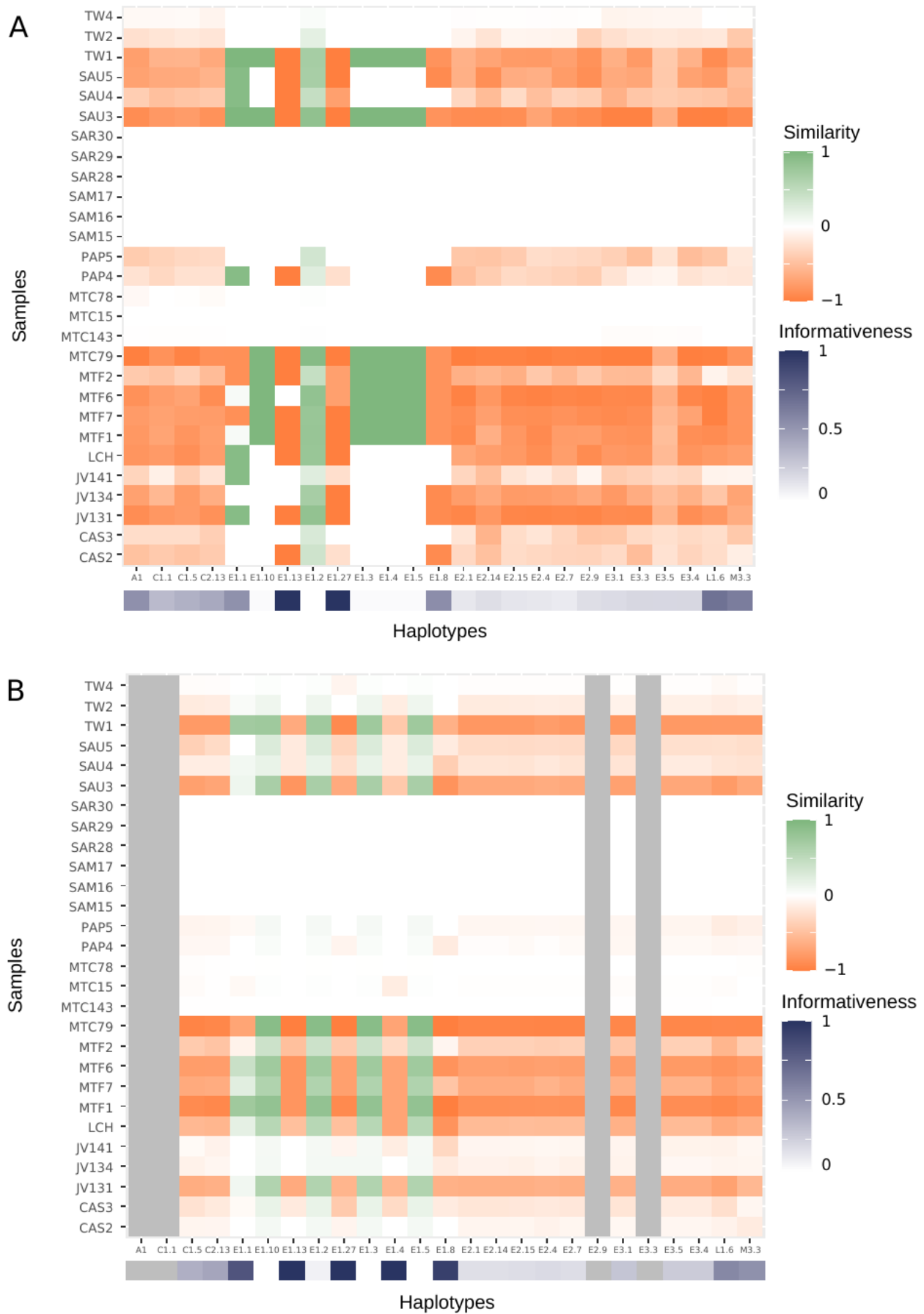


Figure 3. Similarity of ancient samples (Y-axis) to modern haplotypes (X-axis). Similarity score based on plastid (A) and mitochondrial information (B). Low similarity score denotes the presence in the archaeological sample of variants not found in the considered haplotype.

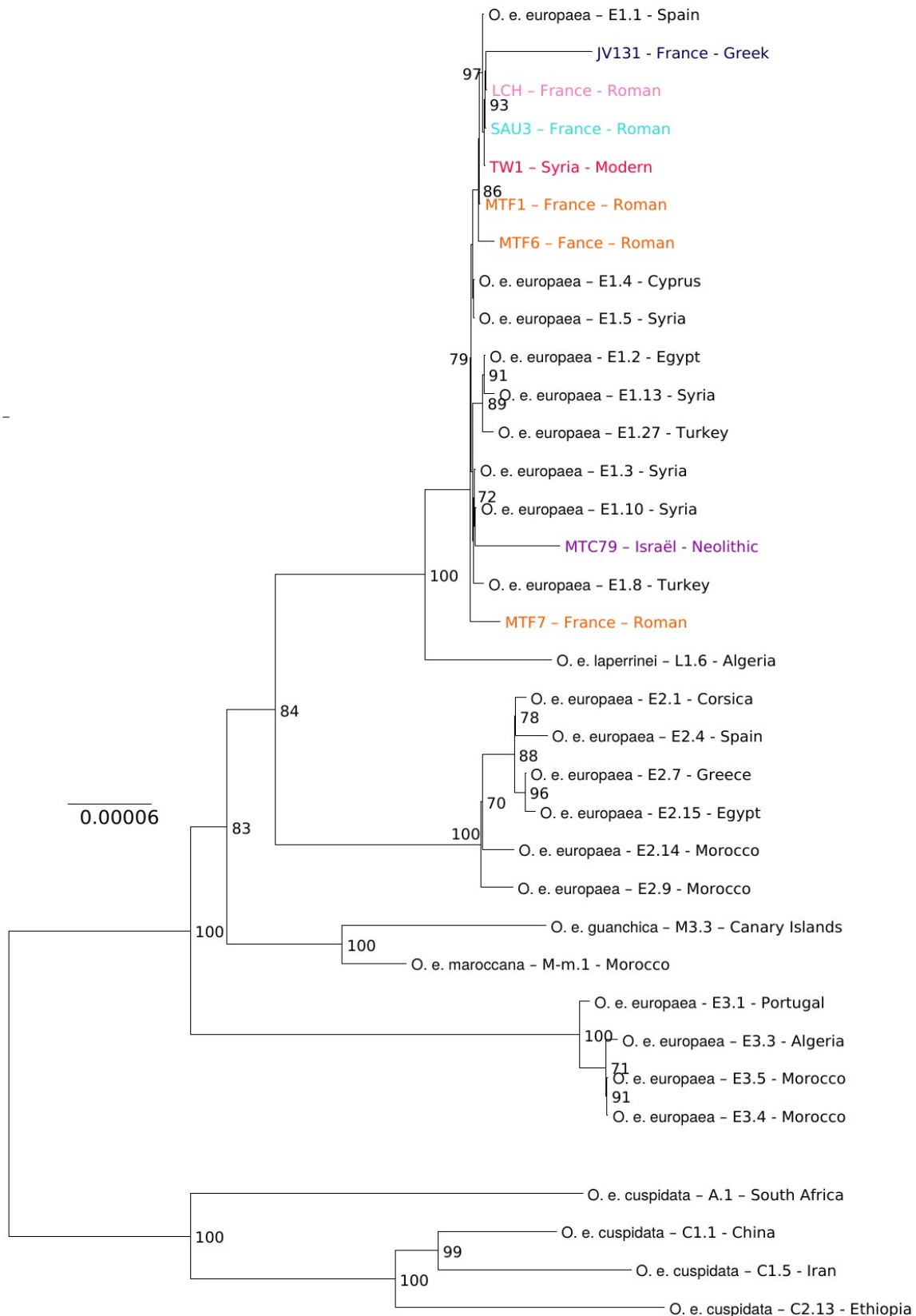
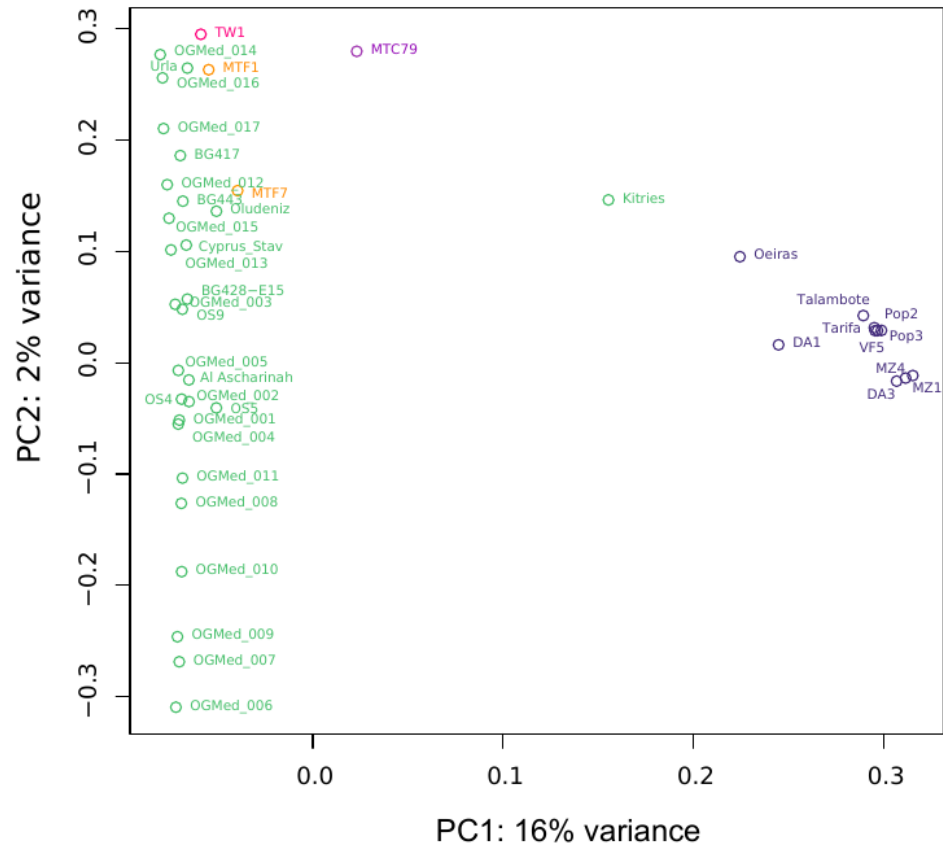


Figure 4. Phylogenetic tree of *O. europaea*, including eight archaeological samples using plastid DNA. Archaeological samples are coloured with the colour code defined in Figure 1.

A



B

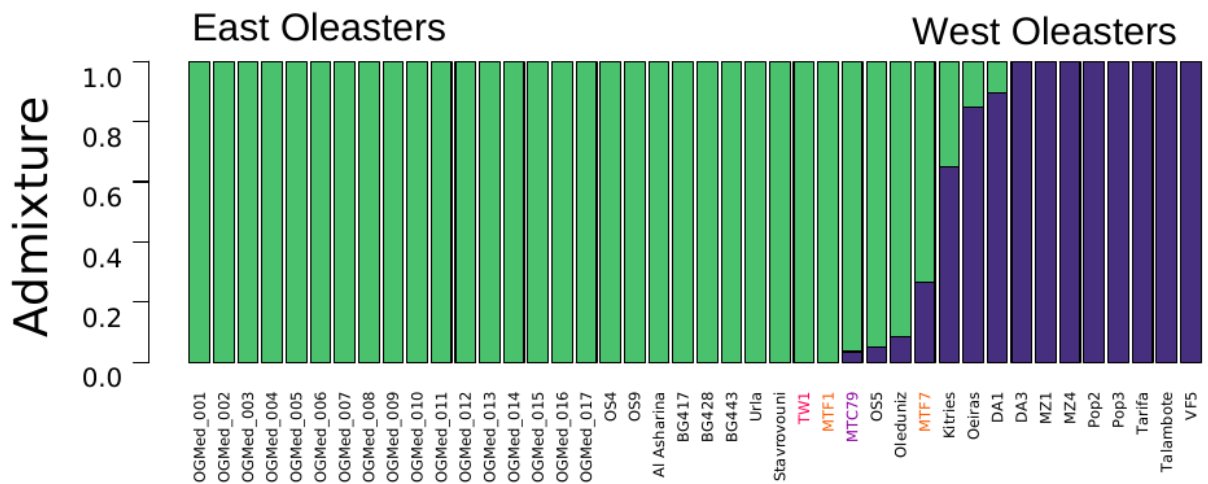


Figure 5. Relationship between archaeological and modern samples based on nuclear SNP data. Only the four archaeological samples with more than $>0.1\times$ coverage were included. Western gene pool is coloured in blue, eastern in green, archaeological samples with the colour code defined in Figure 1. A) PCA plot, B) Clustering result for $K = 2$.

Author contributions

NW and GB designed the study with input from LO, OB, LB, AD, DK, IF, EG, CN, CP, NR, MT, and JFT provided archaeological material. NW and KD did the lab work on archaeological samples. GB and SM did the lab work on modern samples. NW and PR analysed the data with inputs from PAC. PR, NW and GB wrote the paper with inputs from all authors.

Acknowledgments

We are grateful to the Get-Plage sequencing and Genotoul bioinformatics (Bioinfo Genotoul) platforms for sequencing services and providing computing resources.

References

- Angiolillo A, Mencuccini M, Baldoni L. 1999. Olive genetic diversity assessed using amplified polymorphic fragment length polymorphisms. *Theoretical and Applied Genetics* 98: 411–421.
- Belaj A, Muñoz-Diez C, Baldoni L, et al. 2007. Genetic diversity and population structure of wild olives from the north-western Mediterranean assessed by SSR markers. *Annals of Botany* 100: 449–458.
- Besnard G, Khadari B, Villemur P, Bervillé A. 2000. Cytoplasmic male sterility in the olive (*Olea europaea* L.). *Theor. Appl. Genet.* 100:1018–1024.
- Besnard G, Baradat P, Bervillé A. 2001. Genetic relationships in the olive (*Olea europaea* L.) reflect multilocal selection of cultivars. *Theor. Appl. Genet.* 102:251–258.
- Besnard G, Rubio de Casas R, Vargas P. 2007. Plastid and nuclear DNA polymorphism reveals historical processes of isolation and reticulation in the olive tree complex (*Olea europaea*). *Journal of Biogeography* 34:736–752.
- Besnard G, Hernández P, Khadari B, Dorado G, Savolainen V. 2011. Genomic profiling of plastid DNA variation in the Mediterranean olive tree. *BMC Plant Biol.* 11.
- Besnard G, El Bakkali A, Haouane H, Baali-Cherif D, Mukhli A, Khadari B. 2013a. Population genetics of Mediterranean and Saharan olives: geographic patterns of differentiation and evidence for early generations of admixture. *Ann. Bot.* 112:1293–1302.
- Besnard G, Khadari B, Navascués M, Fernández-Mazuecos M, El Bakkali A, Arrigo N, Baali-Cherif D, Brunini-Bronzini de Caraffa V, Santoni S, Vargas P, et al., 2013b. The complex history of the olive tree: from Late Quaternary diversification of Mediterranean lineages to primary domestication in the northern Levant. *Proc. R. Soc. B Biol. Sci.* 280:20122833.
- Besnard G, Terral JF, Cornille A. 2018. On the origins and domestication of the olive: A review and perspectives. *Ann. Bot.* 121:385–403.
- Brändle F, Frühbauer B, Jagannathan M. 2022. Principles and functions of pericentromeric satellite DNA clustering into chromocenters. *Semin. Cell Dev. Biol.* 128:26–39.
- Breton C, Tersac M, Bervillé A. 2006. Genetic diversity and gene flow between the wild olive (oleaster, *Olea europaea* L.) and the olive: several Plio-Pleistocene refuge zones in the Mediterranean basin suggested by simple sequence repeats analysis. *Journal of Biogeography* 34: 1916–1928.

- Contento A, Ceccarelli M, Gelati MT, Maggini F, Baldoni L, Cionini PG. 2002. Diversity of *Olea* genotypes and the origin of cultivated olives. *Theor. Appl. Genet.* 104:1229–1238.
- Da Fonseca RR, Smith BD, Wales N, Cappellini E, Skoglund P, Fumagalli M, Samaniego JA, Carøe C, Ávila-Arcos MC, Huffnagel DE, *et al.* 2015. The origin and evolution of maize in the Southwestern United States. *Nat. Plants* 1:1–5.
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, *et al.* 2011. The variant call format and VCFtools. *Bioinformatics* 27:2156–2158.
- Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, *et al.* 2021. Twelve years of SAMtools and BCFtools. *GigaScience* 10:giab008.
- Diez CM, Trujillo I, Martínez-Urdiroz N, Barranco D, Rallo L, Marfil P, Gaut BS. 2015. Olive domestication and diversification in the Mediterranean Basin. *New Phytol.* 206:436–447.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15–21.
- Drouin G, Daoud H, Xia J. 2008. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. *Mol. Phylogenet. Evol.* 49:827–831.
- Dumolin-Lapègue S, Pemonge M-H, Gielly L, Taberlet P, Petit RJ. 1999. Amplification of oak DNA from ancient and modern wood. *Mol. Ecol.* 8:2137–2140.
- Elbaum R, Melamed-Bessudo C, Boaretto E, Galili E, Lev-Yadun S, Levy AA, Weiner S. 2006. Ancient olive DNA in pits: Preservation, amplification and sequence analysis. *J. Archaeol. Sci.* 33:77–88.
- Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Mol. Ecol.* 14: 2611–2620.
- Galili E, Langgut D, Terral JF, Barazani O, Dag A, Kolska Horwitz L, Ogloblin Ramirez I, Rosen B, Weinstein-Evron M, Chaim S, *et al.* 2021. Early production of table olives at a mid-7th millennium BP submerged site off the Carmel Coast (Israel). *Sci. Rep.* 11:1–15.
- Gros-Balthazard M, Besnard G, Sarah G, Holtz Y, Leclercq J, Santoni S, Wegmann D, Glémin S, Khadari B. 2019. Evolutionary transcriptomics reveals the origins of olives and the genomic changes associated with their domestication. *Plant J.* 100:143–157.
- Gros-Balthazard M, Flowers JM, Hazzouri KM, Ferrand S, Aberlenc F, Sallon S, Purugganan MD. 2021. The genomes of ancient date palms germinated from 2,000 y old seeds. *Proc. Natl. Acad. Sci. U. S. A.* 118:1–10.
- Hansen AJ, Willerslev E, Wiuf C, Mourier T, Arctander P. 2001. Statistical evidence for miscoding lesions in ancient DNA templates. *Mol. Biol. Evol.* 18:262–265.
- Hong-Wa C, Besnard G. 2013. Intricate patterns of phylogenetic relationships in the olive family as inferred from multi-locus plastid and nuclear DNA sequence analyses: A close-up on *Chionanthus* and *Noronhia* (Oleaceae). *Mol. Phylogenet. Evol.* 67:367–378.
- Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L. 2013. MapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29:1682–1684.
- Julca I, Marcet-Houben M, Cruz F, Gómez-Garrido J, Gaut BS, Diez CM, Gut IG, Alioto TS, Vargas P, Gabaldón T. 2020. Genomic evidence for recurrent genetic admixture during the domestication of Mediterranean olive trees (*Olea europaea* L.). *BMC Biol.* 18:1–25.

- Kaniewski D, Van Campo E, Boiy T, Terral JF, Khadari B, Besnard G. 2012. Primary domestication and early uses of the emblematic olive tree: Palaeobotanical, historical and molecular evidence from the Middle East. *Biol. Rev.* 87:885–899.
- Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I. 2015. Clumpak: A program for identifying clustering modes and packaging population structure inferences across *K*. *Molecular Ecology Resources* 15: 1179–1191.
- Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* 15:1–13.
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9:357–359.
- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.
- Lipshitz N, Gophna R, Hartman M, Biger G. 1991. The beginning of olive (*Olea europaea*) cultivation in the Old World: a reassessment. *Journal of Archaeological Science* 18: 441–453.
- Liepert S, Sperisen C, Deguilloux M-F, Petit RJ, Kissling ROY, Spencer M. 2006. Authenticated DNA from ancient wood remains. *Ann. Bot.* 98:1107–1111.
- Mascagni F, Barghini E, Ceccarelli M, Baldoni L, Trapero C, Díez CM, Natali L, Cavallini A, Giordani T. 2022. The singular evolution of *Olea* genome structure. *Front. Plant Sci.* 13:869048.
- Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17:10–12.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20:1297–1303.
- Mascher M, Schuenemann VJ, Davidovich U, Marom N, Himmelbach A, Hübner S, Korol A, David M, Reiter E, Riehl S, et al. 2016. Genomic analysis of 6,000-year-old cultivated grain illuminates the domestication history of barley. *Nat. Genet.* 48:1089–1093.
- Médail F, Quézel P. 2018. Biogéographie de la flore du Sahara: une biodiversité en situation extrême. Conservatoire et Jardin Botanique de Genève, IRD Editions, Marseille.
- Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* 37:1530–1534.
- Novák P, Neumann P, Macas J. 2020. Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. *Nat. Protoc.* 15:3745–3776.
- Oksanen J, Simpson GL, Blanchet FG, Kindt R, Legendre P, Minchin PR, O’Hara RB, Solymos P, Stevens MHH, Szoecs E, et al. 2022. vegan: Community Ecology Package. CRAN-R project.
- Palmer SA, Clapham AJ, Rose P, Freitas FO, Owen BD, Beresford-Jones D, Moore JD, Kitchen JL, Allaby RG. 2012. Archaeogenomic evidence of punctuated genome evolution in *Gossypium*. *Mol. Biol. Evol.* 29:2031–2038.
- Pérez-Escobar OA, Bellot S, Przelomska NAS, Flowers JM, Nesbitt M, Ryan P, Gutaker RM, Gros-Balthazard M, Wells T, Kuhnhauser BG, et al. 2021. Molecular clocks and archeogenomics of

- a late period Egyptian date palm leaf reveal introgression from wild relatives and add timestamps on the domestication. *Mol. Biol. Evol.* 38:4475–4492.
- Pérez-Escobar OA, Tusso S, Przelomska NAS, Wu S, Ryan P, Nesbitt M, Silber M V, Preick M, Fei Z, Hofreiter M, *et al.* 2022. Genome sequencing of up to 6,000-year-old *Citrullus* seeds reveals use of a bitter-fleshed species prior to watermelon domestication. *Mol. Biol. Evol.* 39:1–13.
- Pérez-Zamorano B, Vallebuena-Estrada M, González JM, Cook AG, Montiel R, Vielle-Calzada JP, Delaye L. 2017. Organellar Genomes from a ~5,000-year-old archaeological maize sample are closely related to NB genotype. *Genome Biol. Evol.* 9:904–915.
- Petit RJ, Brewer S, Bordács S, Burg K, Cheddadi R, Coart E, Cottrell J, Csaikl UM, van Dam B, Deans JD, *et al.* 2002. Identification of refugia and post-glacial colonisation routes of European white oaks based on chloroplast DNA and fossil pollen evidence. *For. Ecol. Manage.* 156:49–74.
- Ramos-Madrigal J, Smith BD, Moreno-Mayar JV, Gopalakrishnan S, Ross-Ibarra J, Gilbert MTP, Wales N. 2016. Genome sequence of a 5,310-year-old maize cob provides insights into the early stages of maize domestication. *Curr. Biol.* 26:3195–3201.
- Ramos-Madrigal J, Runge AKW, Bouby L, Lacombe T, Samaniego Castruita JA, Adam-Blondon AF, Figueiral I, Hallavant C, Martínez-Zapater JM, Schaal C, *et al.* 2019. Palaeogenomic insights into the origins of French grapevine diversity. *Nat. Plants* 5:595–603.
- Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* 12:87.
- Schubert M, Ermini L, Sarkissian C Der, Jónsson H, Ginolhac A, Schaefer R, Martin MD, Fernández R, Kircher M, McCue M, *et al.* 2014. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protoc.* 9:1056–1082.
- Staats M, Cuenca A, Richardson JE, Vrieling-van Ginkel R, Petersen G, Seberg O, Bakker FT. 2011. DNA damage in plant herbarium tissue. *PLoS One* 6:e28448.
- Terral JF, Alonso N, Buxó I Capdevila R, Chatti N, Fabre L, Fiorentino G, Marinval P, Jordá GP, Pradat B, Rovira N, *et al.* 2004. Historical biogeography of olive domestication (*Olea europaea* L.) as revealed by geometrical morphometry applied to biological and archaeological material. *J. Biogeogr.* 31:63–77.
- Van de Paer C, Bouchez O, Besnard G. 2018. Prospects on the evolutionary mitogenomics of plants: A case study on the olive family (Oleaceae). *Mol. Ecol. Resour.* 18:407–423.
- Unver T, Wu Z, Sterck L, Turktas M, Lohaus R, Li Z, Yang M, He L, Deng T, Escalante FJ, *et al.* 2017. Genome of wild olive and the evolution of oil biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* 114:E9413–E9422.
- Wagner S, Lagane F, Seguin-Orlando A, Schubert M, Leroy T, Guichoux E, Chancerel E, Bech-Hebelstrup I, Bernard V, Billard C, *et al.* 2018. High-throughput DNA sequencing of ancient wood. *Mol. Ecol.* 27:1138–1154.
- Wales N, Akman M, Watson RHB, Sánchez Barreiro F, Smith BD, Gremillion KJ, Gilbert MTP, Blackman BK. 2019. Ancient DNA reveals the timing and persistence of organellar genetic bottlenecks over 3,000 years of sunflower domestication and improvement. *Evol. Appl.* 12:38–53.

Appendix for Chapter 5

Supporting information includes the following items:

Figure S1. Congruence between plastid and mitochondrial affiliation profiles.

Figure S2. Phylogenetic tree of *O. europaea* including 4 archaeological samples using mitochondrial DNA.

Table S1. Archaeological samples characteristics.

Table S2. Modern samples characteristics.

Table S3. Sequencing statistics for archaeological samples.

Table S4. Statistics of the modern SNPs dataset.

Supplementary Datasets Available at https://github.com/praimondeau/archeogenetics_olive

SD1. Estimates of damage patterns in archaeological samples according to MapDamage.

SD2. Complete genotype information at each sites of our panel for each ancient samples.

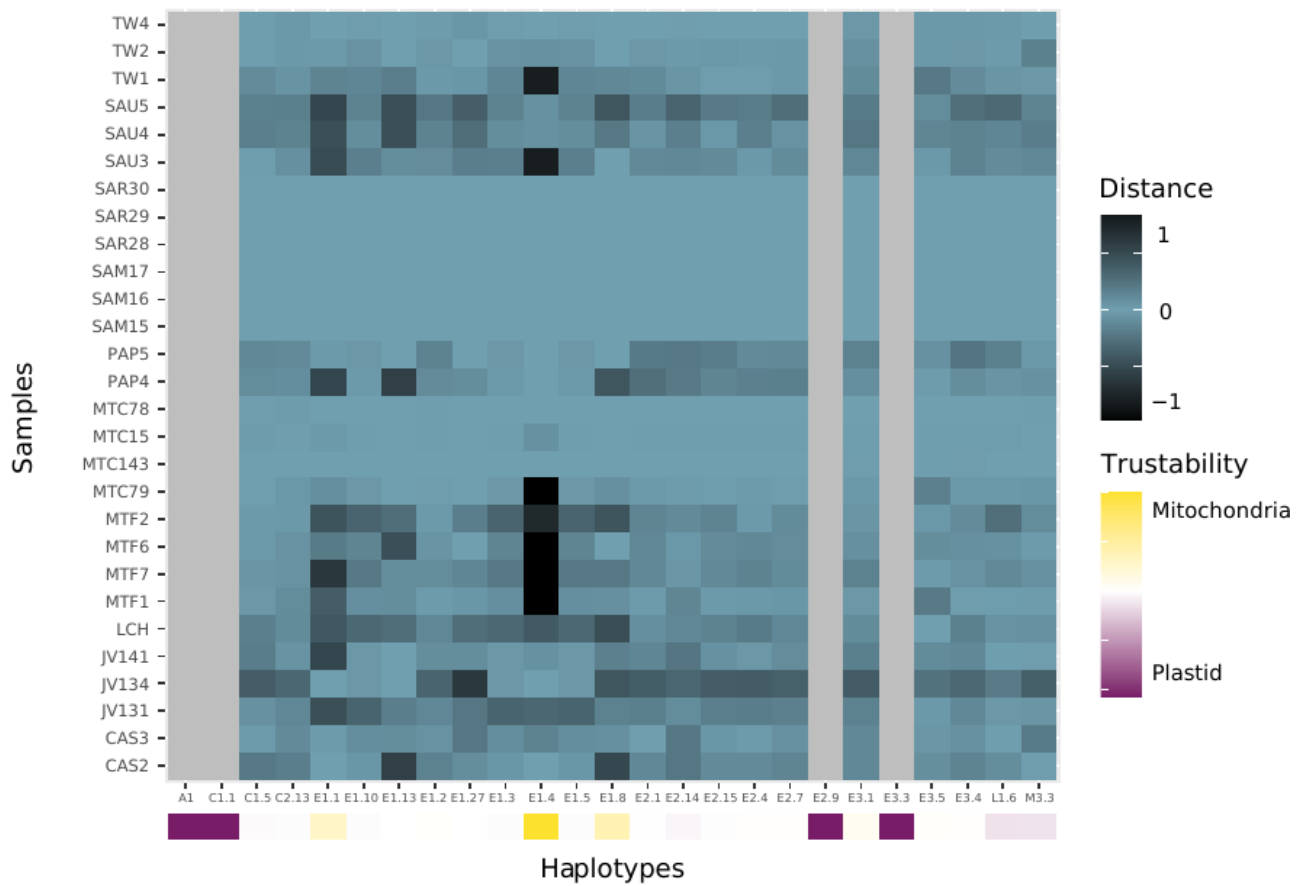


Figure S1. Congruence between plastid and mitochondrial affiliation profiles. Darker blue shades indicates stronger conflicts between mitochondrial and plastid affiliation score. The yellow/purple scale indicates the relative informativeness of organellar genomes for each haplotype. Yellow shades point out haplotypes for which mitochondrial data is more discriminant than plastid one (more haplotype-specific variants).

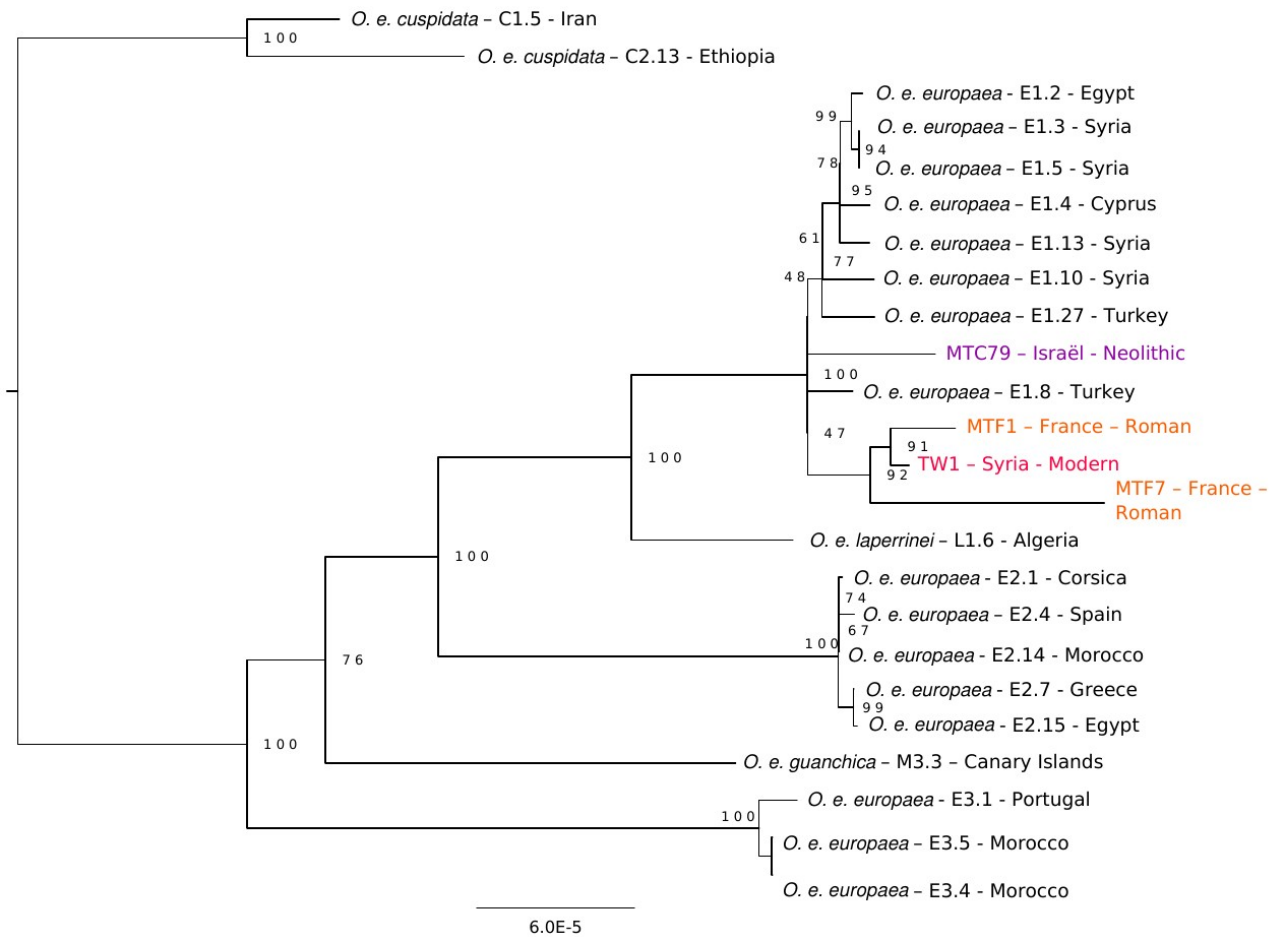


Figure S2. Phylogenetic tree of *O. europaea* including 4 archaeological samples using mitochondrial DNA. The scale is in substitution per site. Ultrafast bootstrap support values are indicated near their respective nodes.

Table S1. Archaeological samples characteristics.

Name	Site	Country	Period/Age	Context	Specimen ID	Type
CAS2	Castelle	France	Roman	PT1021, US1130	2	Olive stone kernel
CAS3	Castelle	France	Roman	PT1021, US1130	3	Olive stone kernel
JV131	JV 13 Jules Verne, Marseille	France	Greek	US02 N°5	1	Olive stone kernel
JV134	JV 13 Jules Verne, Marseille	France	Greek	US02 N°5	4	Olive stone kernel
JV141	JV 14 Jules Verne, Marseille	France	Greek	US01 G3	1	Olive stone kernel
LCH	Lattes-Chenal	France	Roman	US204083		Olive stone kernel
MTF1	Montferrier, Tourbes	France	Roman	PT2052, US2077	1	Olive stone kernel
MTF2	Montferrier, Tourbes	France	Roman	PT2052, US2077	2	Olive stone kernel
MTF6	Montferrier, Tourbes	France	Roman	PT2052, US2077	6	Olive stone kernel
MTF7	Montferrier, Tourbes	France	Roman	PT2052, US2077	7	Olive stone kernel
PAP4	PAP - Pré aux Pêcheurs, Antibes	France	Roman	US3006	4	Olive stone kernel
PAP5	PAP - Pré aux Pêcheurs, Antibes	France	Roman	US3006	5	Olive stone kernel
SAU3	Sauvian, La Lesse	France	Roman	PT3009, US3182	3	Olive stone kernel
SAU4	Sauvian, La Lesse	France	Roman	PT3009, US3183	4	Olive stone kernel
SAU5	Sauvian, La Lesse	France	Roman	PT3009, US3183	5	Olive stone kernel
SAM15	Samos	Greece	2700 BP	Amphora	15	Olive stone kernel
SAM16	Samos	Greece	2700 BP	Amphora	16	Olive stone kernel
SAM17	Samos	Greece	2700 BP	Amphora	17	Olive stone kernel
MTC15	Mt. Carmel	Israel	< 4500 BP	Bag A	15A	Olive stone interior (no visual kernel)
MTC79	Mt. Carmel	Israel	< 4500 BP	Bag A	79A	Olive stone interior (no visual kernel)
MTC78	Mt. Carmel	Israel	< 4500 BP	Bag B	78B	Olive stone interior (no visual kernel)
MTC143	Mt. Carmel	Israel	< 4500 BP	Bag B	143B	Olive stone interior (no visual kernel)
TW1	Tweini TWE10	Syria	Bronze - Iron Age	SG09	1	Olive stone kernel
TW2	Tweini TWE10	Syria	Bronze - Iron Age	SG09	2	Olive stone kernel
TW4	Tweini TWE10	Syria	Bronze - Iron Age	SG09	4	Olive stone kernel
SAR28	Saruq el-Hadid	UAE	Iron Age II	25581-4		Wood
SAR29	Saruq el-Hadid	UAE	Iron Age II	26613-5		Wood
SAR30	Saruq el-Hadid	UAE	Iron Age II	30814-3		Wood

Table S2. Modern samples characteristics. cpDNA = plastid DNA, mtDNA = mitochondrial DNA

Name	Reference	Genepool	Status	Nature	Country	ptDNA haplotype	Technology	cpDNA	mtDNA
Oeiras 3	Van de Paer et al., 2018	West (admixed)	Wild/Feral	WGS	Portugal	E3.1	HiSeq (150 bp)	Y	Y
Vallée du Fango 5	Van de Paer et al., 2018	West	Wild	WGS	Corsica, France	E2.1	HiSeq (150 bp)	Y	Y
Talambote	Generated in this study	West (admixed)	Wild	WGS	Morocco	E2.14	NovaSeq (150 bp)	Y	Y
Tarrifa	Generated in this study	West	Wild	WGS	Spain	E2.4	NovaSeq (150 bp)	Y	Y
Bni Harchen	Generated in this study	West	Wild	WGS	Morocco	E3.5	NovaSeq (150 bp)	Y	Y
Dar Chaoui	Generated in this study	West	Wild	WGS	Morocco	E3.4	NovaSeq (150 bp)	Y	Y
Rajo 10	Generated in this study	East	Wild	WGS	Syria	E1.3	NovaSeq (150 bp)	Y	Y
Rajo 21	Generated in this study	East	Wild	WGS	Syria	E1.5	NovaSeq (150 bp)	Y	Y
Fatrou 9	Generated in this study	East	Wild	WGS	Syria	E1.13	NovaSeq (150 bp)	Y	Y
Al Asharinah 9	Generated in this study	East	Wild	WGS	Syria	E1.10	HiSeq (150 bp)	Y	Y
Urla 6	Generated in this study	East	Wild	WGS	Turkey	E1.8	HiSeq (150 bp)	Y	Y
Kitries 6	Generated in this study	East (admixed)	Wild/Feral	WGS	Greece	E2-7	HiSeq (150 bp)	Y	Y
Oludeniz 1	Generated in this study	East	Wild	WGS	Turkey	E1.27	HiSeq (150 bp)	Y	Y
Stavrovouni C11	Van de Paer et al., 2018	East	Wild	WGS	Cyprus	E1.4	HiSeq (100 bp)	Y	Y
MZ1	Sarah et al., 2017	West	Wild	RNA-seq	Morocco	-	HiSeq (100 bp)	-	-
MZ4	Sarah et al., 2017	West	Wild	RNA-seq	Morocco	-	HiSeq (100 bp)	-	-
DA3	Sarah et al., 2017	West	Wild	RNA-seq	Morocco	-	HiSeq (100 bp)	-	-
DA1	Sarah et al., 2017	West	Wild	RNA-seq	Morocco	-	HiSeq (100 bp)	-	-
OS5	Sarah et al., 2017	East	Wild	RNA-seq	Turkey	-	HiSeq (100 bp)	-	-
OS4	Sarah et al., 2017	East	Wild	RNA-seq	Syria	-	HiSeq (100 bp)	-	-
OGMed_007	Gros-Balthazard et al., 2019	East	Wild	RNA-seq	Turkey	-	HiSeq (100 bp)	-	-
OGMed_006	Gros-Balthazard et al., 2019	East	Wild	RNA-seq	Turkey	-	HiSeq (100 bp)	-	-
OGMed_004	Gros-Balthazard et al., 2019	East	Wild	RNA-seq	Turkey	-	HiSeq (100 bp)	-	-
OS9	Sarah et al., 2017	East	Wild	RNA-seq	Syria	-	HiSeq (100 bp)	-	-
OGMed_011	Gros-Balthazard et al., 2019	East	Wild	RNA-seq	Turkey	-	HiSeq (100 bp)	-	-
OGMed_016	Gros-Balthazard et al., 2019	East	Wild	RNA-seq	Turkey	-	HiSeq (100 bp)	-	-
OGMed_009	Gros-Balthazard et al., 2019	East	Wild	RNA-seq	Turkey	-	HiSeq (100 bp)	-	-
OGMed_008	Gros-Balthazard et al., 2019	East	Wild	RNA-seq	Turkey	-	HiSeq (100 bp)	-	-
OGMed_010	Gros-Balthazard et al., 2019	East	Wild	RNA-seq	Turkey	-	HiSeq (100 bp)	-	-
OGMed_002	Gros-Balthazard et al., 2019	East	Wild	RNA-seq	Turkey	-	HiSeq (100 bp)	-	-
OGMed_001	Gros-Balthazard et al., 2019	East	Wild	RNA-seq	Turkey	-	HiSeq (100 bp)	-	-
OGMed_012	Gros-Balthazard et al., 2019	East	Wild	RNA-seq	Turkey	-	HiSeq (100 bp)	-	-
OGMed_017	Gros-Balthazard et al., 2019	East	Wild	RNA-seq	Turkey	-	HiSeq (100 bp)	-	-
OGMed_005	Gros-Balthazard et al., 2019	East	Wild	RNA-seq	Turkey	-	HiSeq (100 bp)	-	-
OGMed_015	Gros-Balthazard et al., 2019	East	Wild	RNA-seq	Turkey	-	HiSeq (100 bp)	-	-
OGMed_014	Gros-Balthazard et al., 2019	East	Wild	RNA-seq	Turkey	-	HiSeq (100 bp)	-	-

Name	Reference	Genepool	Status	Nature	Country	ptDNA haplotype	Technology	cpDNA	mtDNA
OGMed_013	Gros-Balthazard et al., 2019	East	Wild	RNA-seq	Turkey	-	HiSeq (100 bp)	-	-
OGMed_003	Gros-Balthazard et al., 2019	East	Wild	RNA-seq	Turkey	-	HiSeq (100 bp)	-	-
OB4	Sarah et al., 2017	<i>O. e. cuspidata</i>	Wild	RNA-seq	South Africa	-	HiSeq (100 bp)	-	-
OGCus_001	Gros-Balthazard et al., 2019	<i>O. e. cuspidata</i>	Wild	RNA-seq	South Africa	-	HiSeq (100 bp)	-	-
Iran PV1	Generated in this study	<i>O. e. cuspidata</i>	Wild	WGS	Iran	C1.5	HiSeq (150 bp)	Y	Y
Menagesha 14	Van de Paer et al., 2018	<i>O. e. cuspidata</i>	Wild	WGS	Ethiopia	C2.13	HiSeq (101 bp)	Y	Y
La Gomera 10	Van de Paer et al., 2018	<i>O. e. guanchica</i>	Wild	WGS	Canary Islands	M3.3	HiSeq (150 bp)	Y	Y
Adjelella 10	Van de Paer et al., 2018	<i>O. e. laperrinei</i>	Wild	WGS	Algeria	L1.6	HiSeq (150 bp)	Y	Y
cv. 'Meluki'	Generated in this study	East (admixed)	Cultivated	WGS	Egypt	E1.2	HiSeq (150 bp)	Y	Y
cv. 'Bottegem'	Generated in this study	West (admixed)	Cultivated	WGS	Egypt	E2.15	HiSeq (150 bp)	Y	Y
cv. 'Arbequina'	Generated in this study	East	Cultivated	-	Spain	E1.1		Y	Y
Haut Atlas	Besnard et al., 2011	West		-	Morocco	E2.9		Y	-
Gué de Constantine 20	Besnard et al., 2011	West		-	Algeria	E3.3		Y	-
Guangzhou 1	Besnard et al., 2011	<i>O. e. cuspidata</i>		-	China	C1.1		Y	-
Maui 1	Besnard et al., 2011	<i>O. e. cuspidata</i>		-	Hawaii (South Africa)	A.1		Y	-

Table S3. Sequencing statistics for archaeological samples.

Name	Raw reads	Retained reads	Read length	Clonality	Endogenous content	Nuclear coverage (X)	Cp coverage (X)	Mt coverage (X)
CAS2	5962557	5857064	54.2	4.34%	10.93%	0.025	0.63	0.153
CAS3	29289223	27837487	49.4	4.05%	2.50%	0.024	2.054	0.332
JV131	9369511	8709828	48.4	4.09%	8.17%	0.026	3.886	1.351
JV134	3292773	3212423	58.2	2.52%	4.65%	0.006	1.45	0.143
JV141	4630473	4381520	52.6	7.19%	2.53%	0.004	0.399	0.086
LCH	5348115	5197524	54.6	3.50%	8.09%	0.016	2.883	1.08
MTF1	10974988 SE + 26174395 PE	36179990	54.7	6.56%	21.43%	0.338	18.09	3.843
MTF2	18727977	17649931	47.4	3.44%	5.18%	0.032	0.992	0.595
MTF6	27701970	26442933	48.8	4.55%	9.04%	0.09	4.393	1.969
MTF7	10741197 SE + 57329802 PE	65105662	45.7	5.44%	8.19%	0.195	4.973	1.605
PAP4	6,889,766	6557428	51.3	4.24%	1.87%	0.005	0.528	0.092
PAP5	3,612,285	3529518	59.9	3.14%	3.29%	0.005	0.58	0.154
SAU3	5,120,329	4999616	52.5	4.97%	21.33%	0.047	4.154	1.378
SAU4	6,983,048	6749663	50.6	3.35%	3.04%	0.008	1.012	0.263
SAU5	4,553,834	4352714	47.8	2.99%	6.97%	0.011	1.765	0.439
SAM15	1,808,152	1808152	50.6	49.13%	0.05%	0	0.01	0.001
SAM16	2,281,797	2281797	65.3	27.97%	0.02%	0	0.006	0.001
SAM17	7,306,405	7038585	57.9	41.50%	0.02%	0	0.022	0.001
MTC15	13,669,169	6851488	47.5	5.41%	0.07%	0	0.096	0.014
MTC79	19193922 SE + 331064868 PE	345632305	56.8	8.93%	0.90%	0.113	12.182	4.292
MTC78	5,736,740	13383243	53.7	9.98%	0.09%	0	0.175	0.021
MTC143	7,075,102	5505281	42.6	5.65%	0.10%	0	0.068	0.015
TW1	11,811,617	11625922	48.1	5.58%	38.31%	0.168	13.517	4.125
TW2	7,426,241	7042654	42.5	10.53%	28.26%	0.059	1.021	0.274
TW4	17,921,294	17366872	46.3	5.44%	0.84%	0.005	0.206	0.053
SAR28	7,758,080	7526413	49.1	14.42%	0.02%	0	0.077	0.006
SAR29	5,334,562	5188620	49.5	13.07%	0.01%	0	0.049	0.003
SAR30	10,968,591	10765379	49.3	12.36%	0.01%	0	0.076	0.006
Blank	1,426,399	1,372,744	52.5	41.7%	0.0%	0	0.01	0
Blank	766,315	729,815	51.3	39.9%	0.1%	0	0.01	0

Table S4. Statistics of the modern SNPs dataset.

Samples	Plastome				Mitogenome			
	Covered sites	Filtered SNP	Known SNP	% of panel typed	Covered sites	Filtered SNP	Known SNP	% of panel typed
CAS2	50860	8	0	38.58%	129214	15	0	8.23
CAS3	66573	386	0	31.84%	199518	70	6	17.81
JV131	104944	7	1	84.64%	558326	29	11	61.98
JV134	86559	4	0	68.16%	139094	9	0	8.08
JV141	35656	10	1	25.09%	79109	12	0	7.78
LCH	102490	1	1	81.65%	509653	26	9	56.74
MTF1	122270	1	1	81.65%	736861	39	24	89.22
MTF7	108204	3	2	82.02%	589847	33	11	65.42
MTF6	107409	2	2	86.52%	644386	26	19	77.40
MTF2	59108	7	1	44.94%	344487	18	5	40.72
MTC79	110914	7	1	94.76%	723368	35	23	94.46
MTC143	2736	10	0	0.75%	9626	12	0	0.30
MTC15	2356	14	0	0.00%	7552	13	1	1.35
MTC78	3790	8	0	1.50%	10364	21	0	0.45
PAP4	40637	10	1	25.84%	82607	22	1	5.24
PAP5	49362	7	0	34.83%	130716	15	2	8.08
SAM15	615	4	0	0.00%	1184	1	0	0
SAM16	345	3	0	0.00%	591	2	0	0
SAM17	1064	3	0	0.00%	1330	2	0	0
SAR28	1381	11	0	0.00%	2325	9	0	0
SAR29	1234	15	0	0.00%	1872	5	0	0
SAR30	1462	14	0	0.00%	2426	18	0	0
SAU7	110197	2	2	88.76%	573622	18	9	69.16
SAU8	66943	6	1	46.07%	198953	20	2	17.81
SAU9	89936	1	1	68.54%	295478	29	4	27.69
TW1	111274	2	2	70.79%	709366	31	24	79.49
TW2	37908	0	0	21.72%	146037	4	3	13.02
TW4	10426	16	0	4.49%	36123	27	0	1.95

Conclusions and perspectives

As part of this work, we undertook some of the first genomic-scale studies of the Oleaceae family. This family represents a particularly interesting model system, thanks to the cultural or economic importance of its members, but also due to its evolutionary peculiarities such as variation in reproductive traits.

On polyploidization and the breakdown of heterostyly

In the first chapter, we revised the phylogeny of the family using an exhaustive sampling of all known genera using plastome, mitogemome and nuclear genomic data. This work confirmed the need for taxonomic revision of several genera that were poly- or paraphyletic, a task that is in progress (Dupin *et al.*, 2022; C. Hong-Wa *et al.*, submitted). While we were able to resolve few contentious relationships, our work also highlighted uncertainty to reconstruct the deepest nodes of the paleopolyploid tribe Oleae. Following our work, Dong *et al.* (2022) used nuclear SNP data for a large sampling of Oleaceae in an attempt to characterize the source of these discrepancies. They confirmed that incomplete lineage sorting was misleading phylogenetic reconstruction in Oleae and corroborate our resolution of Schreberineae as sister of all other Oleae. They also suggested Forsythieae and Jasmineae (or a sister ghost lineage) are the parental lineages that produced Oleae as a result of an allopolyploidization event, explaining the unstable placement of these tribes as sister to Oleae.

These findings are of great importance in regards to investigations on the evolution of self-incompatibility systems in the family conducted in chapters 3 and 4. Indeed, it confirms that diploid lineages in Oleaceae exhibit heteromorphic self-incompatibility system, while in the tribe Oleae, that originated from the hybridization of two heteromorphic lineages, the self-incompatibility system is generally homomorphic, with the exception of Schreberineae (C. Hong-Wa *et al.*, submitted). This distribution sustains the hypothetic link between polyploidization and the breakdown of heterostyly in Oleae (Saumitou-Laprade *et al.*, 2010). Based on the distribution of heterostyly in flowering plants, Naiki (2012) confirmed the existence of a correlation between the occurrence of heterostyly and lower ploidy level in some groups but did not conclude there was a direct effect of polyploidization on the breakdown of heterostyly. The dramatic change following hybridization and genome doubling can affect many aspects of a species biology (Qiu *et al.*, 2020) that may in turn promote the transition to different self-incompatibility systems. Whether direct or indirect, the simultaneity of the two phenomena in Oleaceae calls for investigation.

On the homology of heteromorphic and homomorphic diallelic self-incompatibility systems

Self-incompatibility system evolution is undoubtedly the most intriguing question in the family. The discovery of a hemizygous *S*-locus governing homomorphic self-incompatibility in olive brings as many answers as new questions. For instance, it reinforces questions about the homology of homomorphic and heteromorphic di-allelic self-incompatibility systems in Oleaceae (Saumitou-Laprade *et al.*, 2010).

We tried to tackle this question using comparative genomics and in addition to our work in olive, we conducted a transcriptomic study of jasmine heterostyly. We, however, did not identify common gene between differentially expressed genes in jasmine and *S*-locus genes in olive. We also mined available transcriptomic data sets for olive reproductive tissue from cultivars of defined self-incompatibility group, but we did not retrieve any of our candidate genes in them. Self-incompatibility determinants may be triggered specifically during pollination, a factor to keep in mind when designing future study about DSI functional aspects.

The long evolutionary distance between these lineages, in addition to the polyploidization event that separates them, hinder the use of synteny to retrieve homologous candidate regions. Indeed, most olive genes (>50%) are paralogous sequences conserved from the allopolyploidization event (Figure 1A). We did identify a syntenic block including the *S*-locus in olive but investigation at the micro-syntenic scale shown that only the surrounding genes were collinear, and that none of the *S*-locus gene was present within the jasmine region (Figure 1B). It could also be that the reference jasmine genome is not from the morph harboring the hemizygous region. We had similar results with the reference genomes of *Fraxinus excelsior* (known to harbors DSI). For *Osmanthus fragrans* (Oleaceae) and the heterostylous species *Forsythia suspensa* (the floral morph of the sequenced individual is not reported; Li *et al.*, 2020), we did not retrieve any region syntenic to the olive *S*-locus neither and the genes surrounding the *S*-locus are distant of more than 5 Mb on the chromosome with the best synteny with olive chromosome 18 in these assemblies.

Nevertheless, we aggregated nuclear data available in public repository for the family and performed orthology inference with OrthoFinder (Emms and Kelly 2019; see Supplementary Notes for details). Orthogroups were successfully built for two of the *S*-locus genes. They are part of large gene families and exhibit multiple duplication events, but orthologous sequences for the *S*-locus genes were inferred for each of them in *Chrysojasminum fruticans*. For GWHPAOPM038188 ('GIBBERELLIN 2-BETA-DIOXYGENASE 2'), a single ortholog was identified (Figure 2A). From our expression data, it is specifically expressed in *C. fruticans* short-styled individuals. For

GWHPAOPM038191 ('BZR1'), we identified orthologs in *C. fruticans*, *C. bignoniaceum* and *F. suspensa* (Figure 2B). Inspection of expression levels for this transcript revealed it is indeed expressed in *C. bignoniaceum* (homostyle) and in short-styled individuals of *C. fruticans*. We did not detect it in longistyle jasmines (either in *C. fruticans* or *C. odoratissimum*). Both transcripts were too lowly expressed to be significant in our differential expression analyses. Using genome-skimming data generated on two short and long-styled individuals of *C. fruticans*, we confirmed these two genes are specific to short-styled plants. This element suggests that the molecular determinants of homomorphic and heteromorphic self-incompatibility systems might be at least partially common in the family. Homomorphic species would have lost the *S*-locus genes controlling morphological variation but conserved the self-incompatibility features. Investigating cross-species incompatibility reactions between heteromorphic and homomorphic species could provide an answer to this question in case it would demonstrate conservation of compatibility groups.

On the evolution of heterostyly in Oleaceae

If the results presented above provide elements to explain the emergence of the unique homomorphic DSI of Oleaceae, the evolutionary history of heteromorphic systems in the family is itself unclear. Two models have been proposed to explain the emergence of heterostyly. In the first model, a pollen incompatibility mutation first appears in a monomorphic and self-compatible population and get fixed as being advantageous to avoid selfing (Charlesworth and Charlesworth, 1979). Conversely, the first mutation to appear in the second model would impact style length in a population of approach herkogamous flowers (Lloyd and Webb, 1992).

The co-occurrence of stigma-height dimorphism, approach herkogamy and distyly in Jasmineae has been interpreted as evidence for the evolution of distyly through the Lloyd and Webb (1992) model, i.e. approach herkogamous ancestor evolved into a distylous form via stigma-height dimorphism (Ganguly and Barua, 2020). We agree that the occurrence of the three character states, makes *Jasminum* an excellent study system for the evolution of heterostyly. However, heterostyly is also found in other Oleaceae and in the sister lineage Carlemanniaceae. The current hypothesis is therefore that heterostyly is ancestral in the family. The pattern observed in Jasmineae would thus represent loss of heterostyly or reversal. Another alternative hypothesis is that heterostyly evolved independently more than once in the family, as it has been hypothesized in the case of Rubiaceae (Ferrero *et al.*, 2012). Deciphering the molecular determinants of distyly in distinct Oleaceae lineage would probably settle the matter but is a long-reaching goal. Ancestral state reconstruction

could provide a better view on the origin of heterostyly in the family but would still require a greater sampling efforts for this highly diversified lineage (>200 species; Green, 2004).

Again, comparative genomics between species with distinct self-incompatibility status is a powerful approach to examine the different hypotheses on the unique or multiple origins of heterostyly in Oleaceae. Prior to this work, very few genomic data were suitable for this purpose. The mating type of the olive, jasmine and forsythia individuals used to generate the reference genomes are for instance unknown. This unknown parameter may explain the very few orthologous sequences we were able to retrieve for DSI genes. Indeed, given the conservation of the system at least at the scale of the Oleaceae tribe, we would expect to retrieve these genes in other sequenced Oleaceae, unless the sequenced individuals were *ss*.

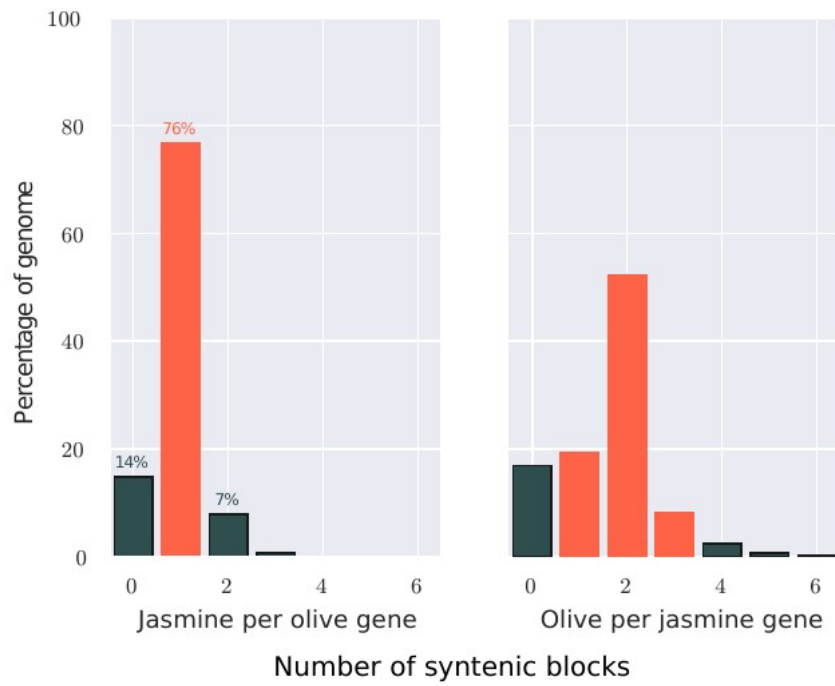
Oleaceae as a model to study the impact of life-history traits on evolutionary rates

Similarly, our investigation of evolutionary rates variation was limited by the availability of suitable data. Indeed, the available genomics datasets were only seldomly accompanied by phenotypic description of traits of interests for evolutionary rates. Mining already generated datasets allowed us to investigate multiple evolutionary questions but also stress the limits of not fit for purpose data. Finding a balance between versatility and precision constitutes a golden rule to successfully navigate the data deluge.

In Chapter 2, we still delve into the remarkable pattern of heterotachy observed in the phylogenetic work we conducted in Chapter 1. We uncovered a potential link between biparental transmissions of plastids and their accelerated evolution. Our hypothesis posits the existence of intraindividual selection of plastids in zygote. Selection among cell lineages is an emerging area of research with cancer genomics at the front end of this exploration. Plants represent an exciting system to study the phenomenon (Watson *et al.*, 2016; Reusch *et al.*, 2021). Indeed, the non-segregation of germline in plants indeed offers the possibilities that somatic mutations get passed through gametes to the next generation (Lanfear, 2018; Hanlon *et al.*, 2019). Somatic mutations may then add to the mutational load of plants and impact evolutionary rates (Schoen and Schultz, 2019). Olive tree could be a particularly interesting model in the matter due to its long-life span and its ability to do clonal propagation (Baali-Cherif and Besnard, 2005).

Our work demonstrates the great potential of genomics to study Oleaceae evolution. As more data is being generated, more questions would be addressable in this remarkably interesting family of non-model angiosperms. We here settled some directions to orientate and facilitate future works.

A



B

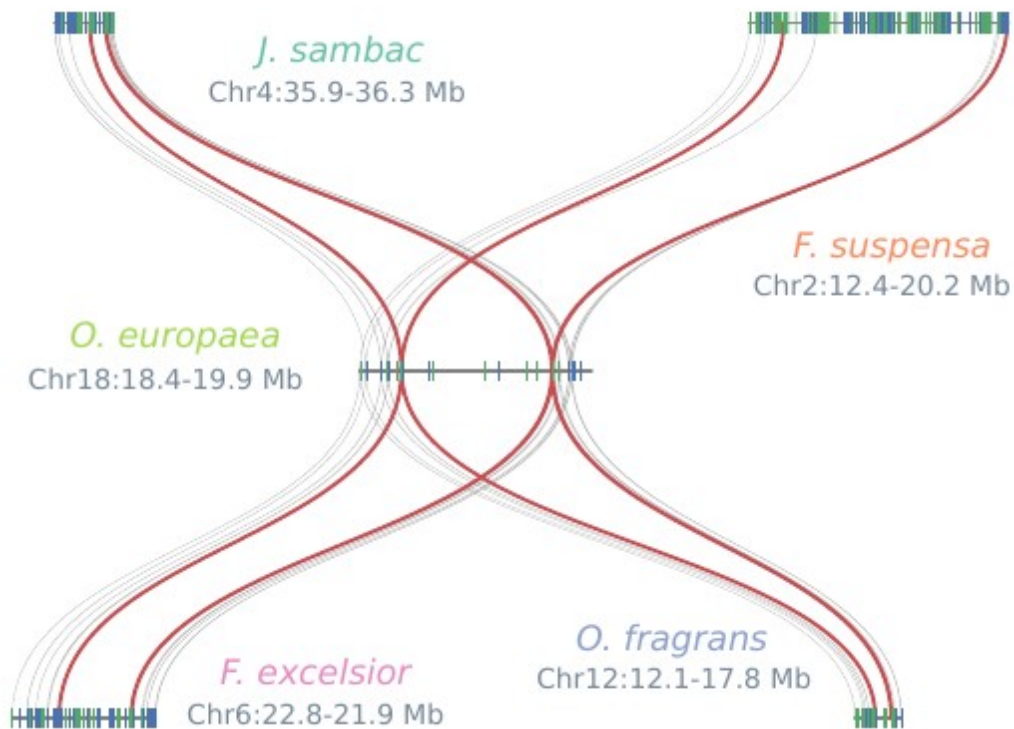
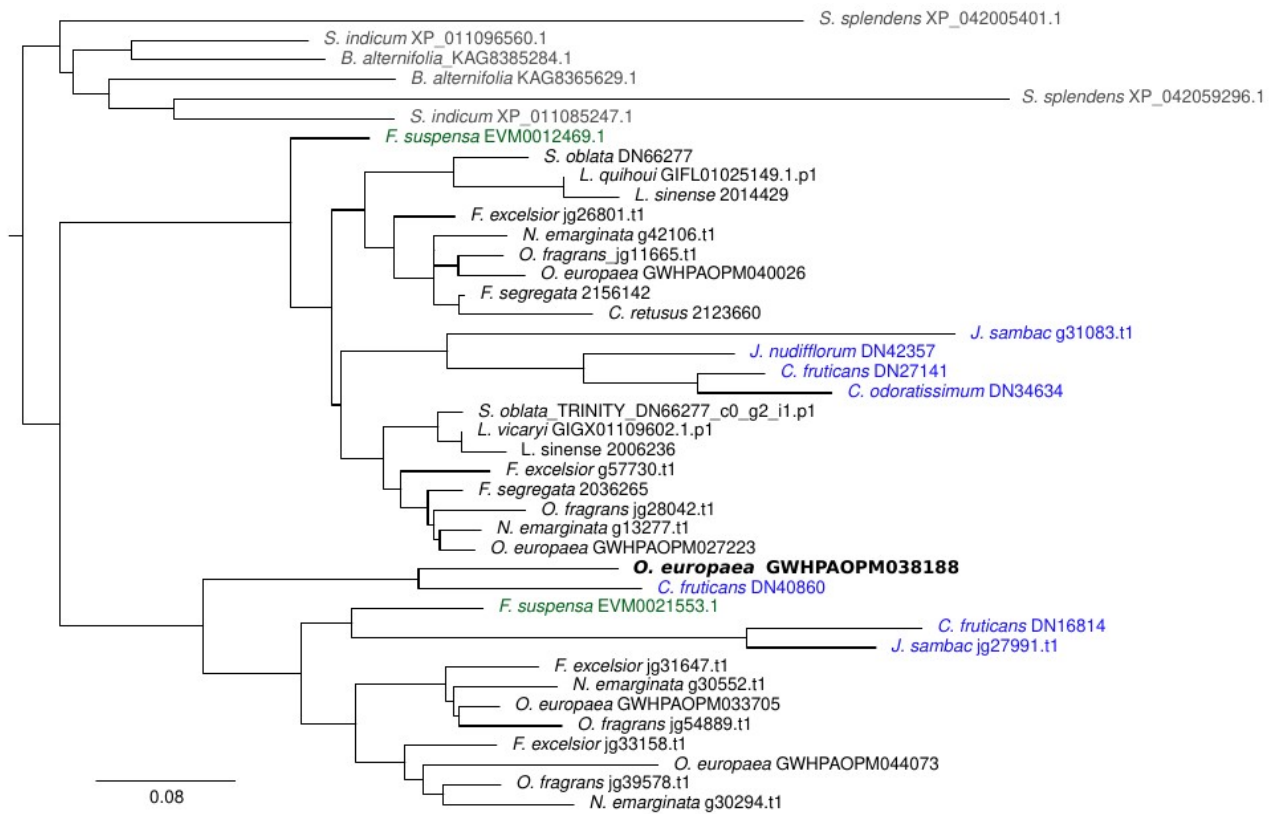
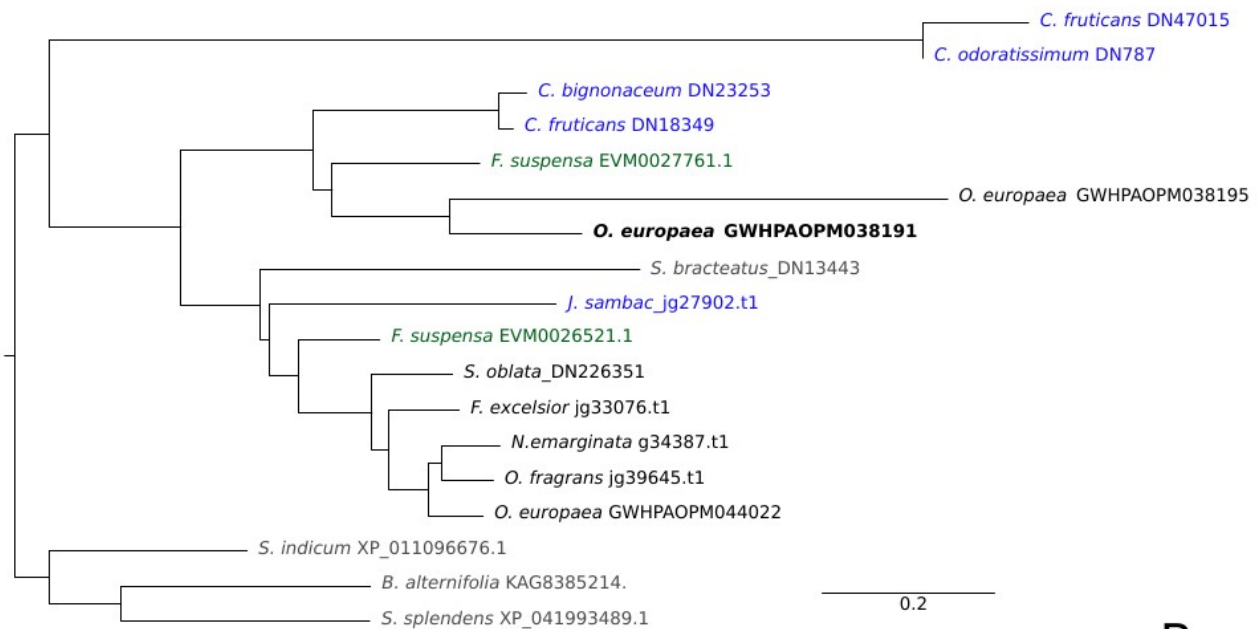


Figure 1. Exploration of Oleaceae synteny. A) Synteny depth between genomes of *O. europaea* (‘Arbequina’) and *J. sambac*. **B)** Microsyntenic relationships with olive *S*-locus. Lines connect syntenic genes, red lines denote the two genes immediately outside the *S*-locus in olive.



A



B

Figure 2. Phylogenetic trees of two multigenic families with members located in the olive *S*-locus A) for GWHPAOPM038188. Paralogs representing duplicates originating before Oleaceae divergence from other Lamiales are not shown. B) for GWHPAOPM038191 and its duplicate in the SI region GWHPAOPM038195. Outgroups are colored in grey, Forsythieae in green, Jasmineae in blue and Oleaceae in black.

Supplementary Notes. Orthology inference

Sampling, sequencing and assembly

We mined available nuclear data for Oleaceae and outgroup species in public repositories (Table 1). We collected eight genomes, as well as seven published transcriptomes. We newly assembled four transcriptomes from publicly available reads and used the three transcriptomes generated in Chapter 4. Finally, we newly sequenced one genome for *Noronhia emarginata*. DNA was extracted from fresh leaves and sequenced on one lane of HiSeq2000. We retrieve RNA-seq reads from SRA for *Jasminum nudiflorum*, *Syringa oblata* and *Silvianthus bracteatus*, a shrubs from Carlemanniaceae, the sister family of Oleaceae. Reads were cleaned with Fastp (Chen *et al.*, 2018). RNA-seq reads were assembled with Trinity (Bryant *et al.*, 2017) with default parameters. Whole-genome sequencing reads for *N. emarginata* were assembled with SOAPdenovo (parameters:k=37, avg_ins=417, map_len=32, pair_num_cutoff=32; (Luo *et al.*, 2012).

Table 1. Details on the nuclear data used for orthology inference.

Taxa	Type	Reference
<i>Arabidopsis thaliana</i>	Genome	GCA_000001735.2
<i>Buddleia alterniflora</i>	Genome	GCA_019426215.1
<i>Chionanthus retusus</i>	Transcriptome	Leebens-Mack et al., 2019
<i>Chrysojasminum bignonaceum</i>	Transcriptome	This study
<i>Chrysojasminum fruticans</i>	Transcriptome	This study
<i>Chrysojasminum odoratissimum</i>	Transcriptome	This study
<i>Forestiera segregata</i>	Transcriptome	Leebens-Mack et al., 2019
<i>Forsythia suspensa</i>	Genome	Li et al. 2020
<i>Fraxinus excelsior</i>	Genome	Sollars et al. 2017
<i>Jasminum nudiflorum</i>	Transcriptome	PRJNA17821
<i>Jasminum sambac</i>	Genome	Xu et al., 2022
<i>Ligustrum quihoui</i>	Transcriptome	PRJNA596355
<i>Ligustrum sinense</i>	Transcriptome	Leebens-Mack et al., 2019
<i>Ligustrum vicaryi</i>	Transcriptome	PRJNA597578
<i>Noronhia emarginata</i>	Genome	This study
<i>Olea europaea</i>	Genome	Rao et al. 2021
<i>Osmanthus fragans</i>	Genome	Yang et al. 2018
<i>Phillyrea angustifolia</i>	Transcriptome	Sarah et al., 2017
<i>Salvia splendens</i>	Genome	GCA_004379255.2
<i>Sesamum indicum</i>	Genome	GCA_000512975.1
<i>Silvianthus bracteatus</i>	Transcriptome	PRJNA636634
<i>Syringa oblata</i>	Transcriptome	PRJNA573685

Genome/transcriptome quality and annotation

Completeness of newly sequenced and annotated genomes and transcriptomes was assessed using BUSCO V4.1.4 with the eudicots odb10 sets (Simão *et al.*, 2015). Genome assemblies were softmasked using Red (Girgis, 2015) and annotated using Braker2 (Brůna *et al.*, 2021) with hints from OrthoDB (Zdobnov *et al.*, 2021) and publicly available RNA-seq data when available. The cleaned reads were mapped against the corresponding genomes using STAR2 (Dobin *et al.*, 2013). We use Transdecoder (Bryant *et al.*, 2017) to predict coding-sequences in transcriptomes using BLASTp hits against the Swissprot database (Altschul *et al.*, 1990) and HMMER search against the Pfam v34 database (Mistry *et al.*, 2021). Finally, protein-coding sequences were extracted from gff annotation files with gffread (Pertea and Pertea, 2020), only keeping complete CDS (no in-frame stop codons/pseudogene, or fully contained sequence). We evaluate annotations completeness by running BUSCO again on the predicted protein sets.

Orthogroups reconstruction

We clustered the protein sequences from 19 Oleaceae and four outgroup species into orthogroups using OrthoFinder v2.5.2 (Emms and Kelly, 2019) with DIAMOND (Buchfink *et al.*, 2021) set in ultra-sensitive mode and the MSA option. The analysis of protein sequences extracted from 10 genome and 12 transcriptome assemblies enable the composition of 40,314 orthogroups, including 90.2% of the total number of proteins analyzed. A species tree was built based on 1,039 of these orthogroups (Figure 1) and was congruent with the current phylogenetic hypothesis for the Oleaceae family (Dupin *et al.*, 2020; Dong *et al.*, 2022).

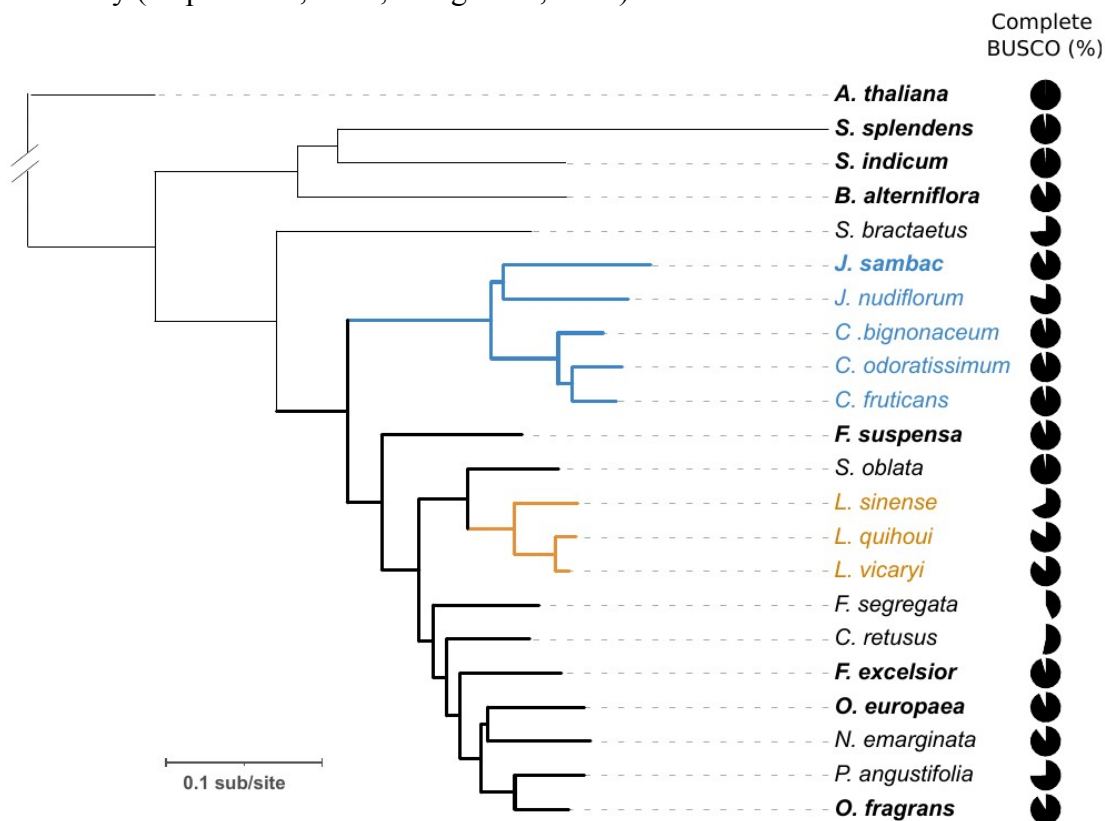


Figure 1. Phylogenetic tree of Oleaceae used for orthogroups reconstruction. Tree was estimated by Orthofinder using 1039 orthogroups with a minimum of 26% of species having single-copy genes in any orthogroup. Thicker lines delimits Oleaceae branches. Bold species names indicates the use of a genome assembly. Orange and blue tips are Core Ligustrinae and Jasmineae, respectively.

References

- Almagro Armenteros JJ, Salvatore M, Emanuelsson O, Winther O, von Heijne G, Elofsson A, Nielsen H. 2019. Detecting sequence signals in targeting peptides using deep learning. *Life Sci. alliance* 2:8.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.
- Baali-Cherif D, Besnard G. 2005. High Genetic Diversity and Clonal Growth in Relict Populations of *Olea europaea* subsp. *laperrinei* (Oleaceae) from Hoggar, Algeria. *Ann. Bot.* 96:823–830.
- Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics Bioinformatics* 3:lqaa108.
- Bryant DM, Johnson K, DiTommaso T, Tickle T, Couger MB, Payzin-Dogru D, Lee TJ, Leigh ND, Kuo T-H, Davis FG, et al. 2017. A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. *Cell Rep.* 18:762–776.
- Buchfink B, Reuter K, Drost H-G. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* 18:366–368.
- Charlesworth B, Charlesworth D. 1979. The maintenance and breakdown of distyly. *Am. Nat.* 114:499–513.
- Chen S, Zhou Y, Chen Y, Gu J. 2018. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34:i884–i890.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15–21.
- Dong W, Li E, Liu Y, Xu C, Wang Y, Liu K, Cui X, Sun J, Suo Z, Zhang Z, et al. 2022. Phylogenomic approaches untangle early divergences and complex diversifications of the olive plant family. *BMC Biol.* 20:92.
- Dupin J, Hong-Wa C, Pillon Y, Besnard G. 2022. From the Mediterranean to the Pacific: re-circumscription towards *Notelaea* s.l. and historical biogeography of a generic complex in Oleinae (Oleaceae). *Bot. J. Linn. Soc.* 200:boac024.
- Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20:238.
- Ferrero V, Rojas D, Vale A, Navarro L. 2012. Delving into the loss of heterostyly in Rubiaceae: Is there a similar trend in tropical and non-tropical climate zones? *Perspect. Plant Ecol. Evol. Syst.* 14:161–167.
- Ganguly S, Barua D. 2020. High herkogamy but low reciprocity characterizes isoplethic populations of *Jasminum malabaricum*, a species with stigma-height dimorphism. *Plant Biol.* 22:899–909.
- Girgis HZ. 2015. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinformatics* 16:227.

- Green PS. Green PS. 2004. Oleaceae. In: Flowering Plants·Dicotyledons. Springer. p. 296–306.
- Hanlon VCT, Otto SP, Aitken SN. 2019. Somatic mutations substantially increase the per-generation mutation rate in the conifer *Picea sitchensis*. *Evol. Lett.* 3:348–358.
- Hong-Wa C, Dupin G, Frasier C, Schatz G, Besnard G. Submitted. Systematics and biogeography of subtribe Schreberinae (Oleaceae), with recircumscription and revision of its Malagasy members.
- Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30:3059–3066.
- Li L-F, Cushman SA, He Y-X, Li Y. 2020. Genome sequencing and population genomics modeling provide insights into the local adaptation of weeping forsythia. *Hortic. Res.* 7:130.
- Lanfear R. 2018. Do plants have a segregated germline? *PLoS Biol.* 16:1–13.
- Lloyd DG, Webb CJ. 1992. The evolution of heterostyly. In: Barrett SCH (ed.), *Evolution and Function of Heterostyly. Monographs on Theoretical and Applied Genetics*, vol 15, pp.151–178, Springer, Berlin, Heidelberg.
- Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Yunjie, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1:18.
- Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al. 2021. Pfam: The protein families database in 2021. *Nucleic Acids Res.* 49:D412–D419.
- Naiki A. 2012. Heterostyly and the possibility of its breakdown by polyploidization. *Plant Species Biol.* 27:3–29.
- Pertea G, Pertea M. 2020. GFF Utilities: GffRead and GffCompare. *F1000Research* 9:1–20.
- Qiu T, Liu Z, Liu B. 2020. The effects of hybridization and genome doubling in plant evolution via allopolyploidy. *Mol. Biol. Rep.* 47:5549–5558.
- Rao G, Zhang J, Liu X, Lin C, Xin H, Xue L, Wang C. 2021. De novo assembly of a new *Olea europaea* genome accession using nanopore sequencing. *Hortic. Res.* 8.
- Revell LJ. 2012. phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* 3:217–223.
- Reusch TBH, Baums IB, Werner B. 2021. Evolution via somatic genetic variation in modular species. *Trends Ecol. Evol.* 36:1083–1092.
- Saumitou-Laprade P, Vernet P, Vassiliadis C, Hoareau Y, De Magny G, Dommée B, Lepart J. 2010. A self-incompatibility system explains high male frequencies in an androdioecious plant. *Science* 327:1648–1650.
- Schoen DJ, Schultz ST. 2019. Somatic Mutation and Evolution in Plants. *Annu. Rev. Ecol. Evol. Syst.* 50:49–73.
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210–3212.

- Sollars ESA, Harper AL, Kelly LJ, Sambles CM, Ramirez-Gonzalez RH, Swarbreck D, Kaithakottil G, Cooper ED, Uauy C, Havlickova L, et al. 2017. Genome sequence and genetic diversity of European ash trees. *Nature* 541:212–216.
- Watson JM, Platzer A, Kazda A, Akimcheva S, Valuchova S, Nizhynska V, Nordborg M, Riha K. 2016. Germline replications and somatic mutation accumulation are independent of vegetative life span in *Arabidopsis*. *Proc. Natl. Acad. Sci. U. S. A.* 113:12226–12231.
- Xu S, Ding Y, Sun J, Zhang Z, Wu Z, Yang T, Shen F, Xue G. 2021. A high-quality genome assembly of *Jasminum sambac* provides insight into floral trait formation and Oleaceae genome evolution. *Mol. Ecol. Resour.*:1–16.
- Yang X, Yue Y, Li H, Ding W, Chen G, Shi T, Chen J, Park MS, Chen F, Wang L. 2018. The chromosome-level quality genome provides insights into the evolution of the biosynthesis genes for aroma compounds of *Osmanthus fragrans*. *Hortic. Res.* 5:72.
- Zdobnov EM, Kuznetsov D, Tegenfeldt F, Manni M, Berkeley M, Kriventseva E V. 2021. OrthoDB in 2020: evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* 49:D389–D393.

Author: Pauline RAIMONDEAU

Title: Mining high-throughput genomic datasets to investigate the evolutionary history of Oleaceae

Supervisor: Guillaume BESNARD

Abstract: Oleaceae is a cosmopolitan plant family gathering several species of interest such as the olive, ash tree, lilac or jasmine. In addition, this family exhibits a large diversity, particularly in traits linked to reproduction. It thus constitutes a great study system, both in terms of fundamental implications to decipher evolutionary processes and economical applications for olive culture notably. In this dissertation, we took advantages of new possibilities offered by the increasing sequencing data to undertake the study of this non-model family evolution at the genomic scale. We investigated several puzzling questions in the family that have been little explored with genomics data. In particular, this work led to the identification of the genomic locus underlying self-incompatibility in olive which brings new insights on the evolution of Oleaceae.

Key-words: genomics, evolution, organellar genome, olive, self-incompatibility

Auteur : Pauline RAIMONDEAU

Titre : Exploration de données génomiques haut-débit pour l'étude de l'histoire évolutive des Oléacées

Directeur de thèse : Guillaume BESNARD

Lieu et date de soutenance :

Résumé : Les Oléacées sont une famille de plantes cosmopolites incluant plusieurs espèces d'intérêt telles que l'olivier, le frêne, le lilas ou le jasmin. Cette famille présente en plus une large diversité, notamment au niveau de traits liés à la reproduction. Elle constitue donc un modèle d'étude particulièrement intéressant autant sur le plan fondamental pour déchiffrer les processus évolutifs, que sur un plan plus appliqué, pour l'oléiculture notamment. Dans le cadre de cette thèse, nous avons tiré profit de la disponibilité croissante de données de séquençage pour entreprendre l'étude de l'évolution de cette famille d'espèces non modèles à l'échelle génomique. Nous avons étudié différentes questions majeures dans l'évolution de cette famille, jusqu'ici peu explorées à l'aide de données génomiques. Ce travail a notamment abouti à l'identification de la région déterminant l'auto-incompatibilité chez l'olivier, apportant de nouvelles informations pour comprendre l'évolution des Oléacées.

Mots-clés : génomique, évolution, organelle, olive, auto-incompatibilité

Discipline administrative : Ecologie - Evolution

Intitulé et adresse du laboratoire :
Laboratoire Ecologie & Diversité Biologique
UMR 5174 (UPS/CNRS/IRD)
Université Toulouse III - Paul Sabatier
118 route de Narbone
31062 Toulouse cedex 9