Universitat Autònoma de Barcelona

# Analysis of the Olive genome

Irene Consuelo Julca Chávez

Centre for Genomic Regulation

Universitat Autònoma de Barcelona

# Analysis of the Olive genome

Irene Consuelo Julca Chávez

Universitat Autònoma de Barcelona
Faculty of Bioscience
Department of Animal Biology, Plant Biology and Ecology

TESI DOCTORAL UAB 2017

# Analysis of the Olive genome

A thesis submitted by Irene Consuelo Julca Chávez for the degree of Doctor in Plant Biology and Biotechnology under the direction of Dr. Toni Gabaldón, Dr. Pablo Vargas, and tutored by Dr. Josep Alluè

This thesis has been inscribed in the Plant Biology and Biotechnology PhD program from the Universitat Autònoma de Barcelona

**Director**
**Dr. Toni Gabaldón**
Centre for Genomic Regulation

**Codirector**
**Dr. Pablo Vargas**
Royal Botanical Garden of Madrid

**Tutor**
**Dr. Josep Alluè**
Universitat Autònoma de Barcelona

**PhD student**
**Irene Julca**
Universitat Autònoma de Barcelona

Barcelona, September 2017

*A mis padres*
*Delia Chávez y José Julca.*

*"Discovery consists of seeing what everybody has seen,*
*and thinking what nobody has thought."*

*Albert Szent-Györgyi*

# Acknowledgments

A veces pequeñas decisiones cambian el curso de nuestras vidas. Todo empezó cuando decidí dejar el laboratorio y las pipetas por bioinformática. Que idea más loca, iniciar todo de nuevo y encima trabajando en genomas de plantas!

En esta historia hubo un poco de suerte porque encontré un grupo muy agradable, aquí en Barcelona. En este pequeño punto en la tierra muchas cosas de mi vida han cambiado, incluso algunas ideas. Aún recuerdo mi primera entrevista con Toni, que nervios y luego todo calma en un lugar que parecía más que un centro de investigación un hotel con vista al mar. No puedo mentir, la ubicación del centro fue una de las primeras cosas que me impresionaron. Desde esa primera entrevista, yo supe que Toni era una persona sencilla y amable y que yo quería trabajar con él. Muchas gracias Toni por toda la paciencia que has tenido conmigo durante todo este período. Gracias por siempre ayudarme a ver el lado positivo de los resultados, una cosa importante en esta carrera.

Luego conocí a Pablo, una persona muy entusiasta, que a pesar de la distancia, siempre ha encontrado la manera de estar en constante comunicación, sobre todo en los momentos claves. Gracias por guiarme cuando yo estaba perdida en taxonomía y por incentivarme a estudiar filogenias. Gracias por las instructivas charlas durante nuestras largas llamadas telefónicas y por contestarme siempre, incluso cuando te he escrito tarde.

Durante estos años, una persona ha estado guiándome desde el comienzo. Nada hubiera sido igual sin su apoyo. Estoy segura que he tenido mucha suerte, no todos inician con una guia como Marina. Muchas gracias por enseñarme todas las herramientas que finalmente he utilizado, espero bien.

Gracias, por todo el tiempo que te has tomado para ayudarme durante este periodo y también por las fascinantes discusiones sobre los extraños resultados que íbamos encontrando a lo largo de estos años.

Each day from the beginning to the end of this period I was surrounded by amazing people. I want to thank you all, for every moment that we've shared, even the ones that I can't remember now. Miguel, always on time for the breakfast, the lunch, and the coffee break. Gracias por compartir tus ideas, a veces locas, pero siempre interesantes. Cinta, siempre amable y sonriente, gracias por todos los tips sobre el otro mundo de la genética de poblaciones. Ernst, my 'friendly' neighbor, thanks for helping me to solve some problems with my computer and for your valuable advices about statistics. Laia, por la compañía en la parte final de este período y también, por alquilar los lugares 'más baratos' para nuestros fabulosos retreats. Salva, that when was in the group was always solving my computer problems, of course they were never my fault. Alex, for your enjoyable company during the first years of this period. The girls of the wet lab: Ester, Susana and Ewa, was always nice to share a moment with you. Verónica, 'obrigado' for your funny company, and for the olive picture that you found. Hrant, for the nice company in the dark side of the lab. Jesse, to always have the correct pronunciation of the English words. Fran because was almost always available to go out. Pedro, Rosa, and Uciel for the nice advices. Damian, who has very 'interesting' ideas, thanks for helping me with the olive at the beginning. I want to thanks also the people that stayed in the lab for a short period, but have left nice memories: Houda and Radina.

Thanks to the nice people with whom I've shared many adventures inside and outside the CRG. Davide, who always was on time, ready for all, from dinners out to a walk in the mountains. Silvina, for your modernist advices. Mimma for all the countless dinners that we've shared. Alessandra for all the room escapes that we have won together. To my friends that in some moments of this period had shared time with me. Diana, gracias por tus concejos y por las bonitas escapadas juntas. Patty, por compartir tus locuras conmigo. Rosany y Meltem, por ser mis primeras guías en Barcelona.

To the people that was present before and from whom I've learned a lot: Mercedes, Maximino and María.

Charlotte and Josep, who made easier all the bureaucratic process along all these years.

The security training office because, without knowing, allowed me to meet a person that becomes an important part of this history. My first friend at the CRG, Riccardo. Thank you for being the accomplice of my crazy ideas. All these years will be not the same without you. Thank you very much for your patience during this period, overall the last part of it. I know that probably you wanted to run away when I was writing the thesis, but still you were here, smiling and always giving me the strength to continue.

Pancrazio e Carmen, con cui ho condiviso molti bei momenti. Grazie per aver reso i miei giorni piú belli con la vostra presenza. Per essere cosí ospitali e per avermi fatto sentire parte della vostra famiglia.

The most important part of a history is the end. Quiero agradecer a mi familia. Mis tíos José y Vilma, quienes han estado presentes siempre que los he necesitado. Rita, que desde que tengo memoria ha estado cuidándonos. Mis hermanos, José Luis y Carlos, que siempre están allí para escucharme. Mis padres José y Delia por su incondicional amor y apoyo. José que siempre me da fuerzas y ánimo para terminar las cosas que inicio. Delia, una persona que desde la distancia sabe guiarme y que sus consejos cambiaron esta historia. Finalmente a ti, que sin ti nada sería posible.

Thank you all!

Irene Consuelo Julca Chávez.

Barcelona, September 2017.

# Abstract

The olive tree (*Olea europaea*, Oleaceae) is an iconic plant of Mediterranean countries for cultural, historical and biological reasons. The olive species comprises six subspecies (*europaea*, *maroccana*, *cerasiformis*, *laperrinei*, *guanchica*, and *cuspidata*) that together form the so-called *O. europaea* complex. Likewise, the subsp. *europaea* is divided into two taxonomic varieties: var. *europaea*, that comprises all the cultivated forms, and var. *sylvestris* (also called oleaster), that includes the wild forms. The olive tree has been intensively cultivated since 6,000 years ago, coinciding with the emergence of early Mediterranean civilizations. Because of the interest of the drupes both as table olives and as raw material to produce olive oil, the olive tree is an essential crop across the Mediterranean basin. This doctoral thesis aims to provide insights into the biology and the evolution of the cultivated olive and relatives. To this end, we sequenced, assembled, and annotated a reference genome obtained from a single individual (*O. europaea* L. var. *europaea*). Phylogenomic analysis and assessment of allelic relative coverage suggest up to four polyploidization events in the evolutionary history of the olive. Two ancient allopolyploidization events at the base of the family Oleaceae (Eocene-Late Cretaceous), and the tribe Oleeae (Oligocene-Miocene), followed by two polyploidizations in the ancestor of *O. europaea* (Miocene-Pliocene) since its divergence from *Phillyrea angustifolia*. In order to study the diversity and phylogenetic relationships in the *O. europaea* complex, we additionally sequenced the genome of at least one individual per subspecies. Our results show that cultivated olive trees exhibit less nucleotide diversity when compared with wild relatives. Different sets of genes were found to be under positive selection in each cultivar included in this study ('Arbequina', 'Beladi', 'Farga', 'Picual', 'Sorani'). In addition to hybridization involving polyploidization (allopolyploidization), phylogenomic analysis revealed extensive homoploid hybridization among lineages of the *O. europaea* complex, which results in a continuous gene flow from wild to domesticated olive trees. In particular, cv. ‘Farga’ has a different origin than the other cultivars included in this study, and shows evidence for secondary domestication events in the Iberian Peninsula. In summary, this study helps unravel the evolutionary history of *O. europaea*, and uncover a complex scenario of polyploidization and hybridization that resulted in recurrent gene duplications.

# Resumen

El olivo (*Olea europaea*, Oleaceae) es una planta icónica en el Mediterráneo por razones culturales, históricas y biológicas. El olivo como especie está formado por seis subespecies (*europaea*, *maroccana*, *cerasiformis*, *laperrinei*, *guanchica*, y *cuspidata*) que juntas forman el llamado complejo *O. europaea*. Del mismo modo, la subsp. *europaea* se divide en dos variedades: var. *europaea*, que comprende las formas cultivadas, y var. *sylvestris* (también llamado oleaster), que incluye las formas silvestres del Mediterráneo. El olivo ha sido cultivado intensivamente desde hace aproximadamente 6,000 años, coincidiendo con la emergencia de civilizaciones tempranas en el Mediterráneo. Debido al gran interés en sus frutos, como aceitunas de mesa y como material para aceite de oliva, el olivo es considerado un cultivo esencial en la cuenca Mediterránea. Esta tesis doctoral tiene como objetivo aportar conocimientos sobre la biología y la evolución de los olivos cultivados y linajes cercanos. Con este fin, secuenciamos, ensamblamos y anotamos un genoma de referencia correspondiente a un único individuo (*O. europaea* L. var. *europaea*). Análisis filogenómicos y evaluaciones del coverage relativo de alelos sugieren que en la historia evolutiva del olivo ocurrieron un mínimo de cuatro poliploidizaciones. Dos alopoliploidizaciones localizadas en la base de la familia Oleaceae (Eoceno - Cretácico tardío) y en la base de la tribu Oleeae; seguidas de dos poliploidizaciones en el ancestro de *O. europaea* (Mioceno-Plioceno) luego de su divergencia de *Phillyrea angustifolia*. Con el objetivo de estudiar la diversidad y las relaciones filogenéticas en el complejo *O. europaea*, secuenciamos adicionalmente el genoma de al menos un individuo por cada subespecie. Nuestros resultados muestran que los olivos cultivados tienen menos diversidad nucleotídica cuando son comparados con los linajes silvestres. Diferentes genes están bajo selección positiva en cada cultivariedad incluida en este estudio ('Arbequina', 'Beladi', 'Farga', 'Picual', 'Sorani'). Además de hibridación que involucra poliploidización, los análisis filogenómicos revelaron extensivos procesos de hibridazación homoploide

entre los lineajes del complejo *O. europaea*, que resulta en un continuo flujo genético desde olivos silvestres hacia olivos domesticados. En particular, el cv. 'Farga' tiene un origen diferente a las otras cultivariedades incluidas en este estudio y aporta evidencia de domesticación secundaria en la península Ibérica. En resumen, este estudio permite entender la historia evolutiva de *O. europaea*, y descubre un complejo escenario de poliploidizaciones e hibridaciones que han resultado en duplicaciones génicas recurrentes.

# Contents

**Acknowledgments** ix

**Abstract** xiii

**Resumen** xv

**Contents** xx

**1 Introduction** 1

1.1 Systematics and evolution of the olive tree . . . . . . . . . . . . 4

1.1.1 Order Lamiales and family Oleaceae . . . . . . . . . . . 5

1.1.2 The *Olea europaea* complex . . . . . . . . . . . . . . . . . 8

1.2 Domestication of the olive tree . . . . . . . . . . . . . . . . . . . 13

1.2.1 History . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14

1.3 Hybridization . . . . . . . . . . . . . . . . . . . . . . . . . . . 15

1.3.1 Hybridization and *O. europaea* . . . . . . . . . . . . . . 17

1.3.2 Methods to identify hybrids . . . . . . . . . . . . . . . . 17

1.4 Polyploidization . . . . . . . . . . . . . . . . . . . . . . . . . . 19

1.4.1 Autopolyploidization and allopolyploidization . . . . 23

1.4.2 Polyploidy and domestication . . . . . . . . . . . . . . 25

1.4.3 Diploidization . . . . . . . . . . . . . . . . . . . . . . . . 27

1.4.4 Olive tree and polyploidy . . . . . . . . . . . . . . . . . 28

1.5 Methods to estimate polyploidy . . . . . . . . . . . . . . . . . . 29

1.5.1 Chromosome number and nuclear DNA content . . . . 29

1.5.2 Synteny . . . . . . . . . . . . . . . . . . . . . . . . . . . 31

1.5.3 Phylogenomic methods . . . . . . . . . . . . . . . . . . 31

1.6 Final remarks . . . . . . . . . . . . . . . . . . . . . . . . . . . 37

**Hypothesis** 39

**Objectives** 41

**2 Olive genome** 43

2.1 Abstract . . . . . . . . . . . . . . . . . . . . . . . . . . . . 47

2.2 Data description . . . . . . . . . . . . . . . . . . . . . . . . 47

2.2.1 Sequencing . . . . . . . . . . . . . . . . . . . . . . . 47

2.2.2 Genome assembly . . . . . . . . . . . . . . . . . . . . 51

2.2.3 Partial assembly of an olive tree associated fungus: *Aureobasisium pullulans* . . . . . . . . . . . . . . . . . . . . 59

2.2.4 Olive tree genome annotation . . . . . . . . . . . . . . . 60

**3 Phylogenomics of the olive tree** 69

3.1 Abstract . . . . . . . . . . . . . . . . . . . . . . . . . . . . 73

3.2 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . 73

3.3 Results and discussion . . . . . . . . . . . . . . . . . . . . . 76

3.3.1 Gene order analysis indicates multiple polyploidizations in the Lamiales . . . . . . . . . . . . . . . . . . . . 76

3.3.2 The olive phylome . . . . . . . . . . . . . . . . . . . . 76

3.3.3 Phylogenetic analysis reveals an ancient allopolyploidization in Lamiales . . . . . . . . . . . . . . . . . . 79

3.3.4 Increased phylogenetic resolution provided by transcriptomes uncovers allopolyploidization at the base of the tribe Oleeae . . . . . . . . . . . . . . . . . . . . 81

3.4 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . 86

3.5 Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . 87

3.5.1 Gene order analysis . . . . . . . . . . . . . . . . . . . 87

3.5.2 Phylome reconstruction . . . . . . . . . . . . . . . . . 88

3.5.3 Incorporation of transcriptomic data in the olive phylome . . . . . . . . . . . . . . . . . . . . . . . . . . 88

3.5.4 Species tree reconstruction . . . . . . . . . . . . . . . 89

3.5.5 Detection and mapping of orthologs and paralogs . . . 90

3.5.6 GO term enrichment . . . . . . . . . . . . . . . . . . . . . 90
3.5.7 Topological analysis . . . . . . . . . . . . . . . . . . . . 91
3.5.8 Transversion rate at fourfold degenerate sites (4DTv) . 92
3.5.9 Divergence times . . . . . . . . . . . . . . . . . . . . . 92
3.A Supplementary material . . . . . . . . . . . . . . . . . . . . 93

**4 Genome analysis of the *Olea europaea* complex 103**
4.1 Abstract . . . . . . . . . . . . . . . . . . . . . . . . . . . 107
4.2 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . 108
4.3 Results and discussion . . . . . . . . . . . . . . . . . . . . 110
4.3.1 Patterns of genetic polymorphism in *O. europaea* . . . . 110
4.3.2 Identification of genes selected in the domestication of olives . . . . . . . . . . . . . . . . . . . . . . . . . 113
4.3.3 Patterns of allelic representation in heterozygous positions suggest basal tetraploidy . . . . . . . . . . . . . 119
4.3.4 Phylogenetic relationships in the *O. europaea* complex . 123
4.3.5 Concluding remarks . . . . . . . . . . . . . . . . . . . . 130
4.4 Material and Methods . . . . . . . . . . . . . . . . . . . . . 130
4.4.1 Genome sequences . . . . . . . . . . . . . . . . . . . . . 130
4.4.2 Assembly of the plastid and mitochondrial genomes . 131
4.4.3 Detection of Single Nucleotide Polymorphisms . . . . 132
4.4.4 SNPs characterization . . . . . . . . . . . . . . . . . . . 133
4.4.5 Selection Tests . . . . . . . . . . . . . . . . . . . . . . . 133
4.4.6 Analysis of K-mer Spectra in the Reference Genome Sequence . . . . . . . . . . . . . . . . . . . . . . . . . 134
4.4.7 Ploidy estimation . . . . . . . . . . . . . . . . . . . . . 134
4.4.8 Admixture Mapping . . . . . . . . . . . . . . . . . . . . 135
4.4.9 Phylogenetic analysis . . . . . . . . . . . . . . . . . . . 136
4.A Supplementary material . . . . . . . . . . . . . . . . . . . . 137

**5 Summarizing Discussion 147**
5.1 Olive genome . . . . . . . . . . . . . . . . . . . . . . . . . . 149
5.2 Polyploidy in *Olea europaea* . . . . . . . . . . . . . . . . . . 153

5.3 Hybridization of the *O. europaea* complex . . . . . . . . . . . . 157
5.4 Origin of cv. 'Farga' . . . . . . . . . . . . . . . . . . . . . . . . 160
5.5 Gene selection in olive . . . . . . . . . . . . . . . . . . . . . . . 160
5.6 Future perspectives . . . . . . . . . . . . . . . . . . . . . . . . . 161

**6 Conclusions 163**

**7 Appendix: List of publications 167**

**References 171**

# 1 Introduction

## Introduction

Domestication and polyploidy (also known as whole genome duplication - WGD) are key processes in plant evolution that are not completely independent from each other. Polyploids can show many traits that are absent in their diploid progenitors. Some of these traits, such as a higher genetic diversity, mutational robustness, heterozygosity, and heterosis make polyploids suitable material for domestication and breeding (Renny-Byfield and Wendel, 2014). The advent of genome sequencing and comparative mapping studies in plants has uncovered many polyploidization events in the history of a growing number of crops (Jarvis et al., 2017; Montero-Pau et al., 2017; Badouin et al., 2017; Zhang et al., 2015b). These include, among many others, recent polyploidizations in maize, wheat, canola, or banana (D'Hont et al., 2012; Chalhoub et al., 2014; International Wheat Genome Sequencing Consortium, 2014; Messing, 2009). Whether ploypoidization in crops predates or follows domestication has been a matter of discussion (Fang and Morrell, 2016; Salman-Minkov et al., 2016). Recently, it has been proposed that polyploid plants are more likely to be domesticated than their wild diploid relatives, implying that the most likely order of events is that domestication follows polyploidy, rather than the other way around (Salman-Minkov et al., 2016).

The Mediterranean olive tree, *Olea europaea* subsp. *europaea* var. *europaea*, is one of the most ancient cultivated fruit trees in the Mediterranean basin. Its domestication dates back to the Early Bronze Age (6,000 years ago), and the breeding and dispersion of this crop has been tightly linked to the history of Mediterranean civilizations (Zohary and Spiegel-Roy, 1975; Besnard et al., 2013b; Kaniewski et al., 2012). The main plant part selected from wild relatives of the Mediterranean olive tree is the fruit, either as directly edible fruits or as a source for oil (Zohary and Spiegel-Roy, 1975; Kaniewski et al., 2012). The specific place of initial domestication has been a matter of discussions, and the question whether cultivated varieties originated from a single domestication event or from several parallel events, is still debated (Besnard and Rubio de Casas, 2016; Díez and Gaut, 2016).

*Olea europaea* L. belongs to the order Lamiales and the family Oleaceae.

It is further classified into six different subspecies (*europaea*, *maroccana*, *cerasiformis*, *laperrinei*, *guanchica*, and *cuspidata*) which together form the *O. europaea* complex. In the subsp. *europaea*, two varieties are recognized: var. *sylvestris*, that comprises the wild forms of the olive trees; and var. *europaea*, that comprises the cultivated forms (Green, 2002; Vargas et al., 2000). These individuals have an allogamous mode of reproduction and some of them are self-incompatible or male-sterile (Besnard et al., 2000; Breton et al., 2017). This is one of the reasons why the cultivated olive trees have historically been propagated vegetatively, either by cuttings or grafts (Zohary and Spiegel-Roy, 1975).

Spain, is one of the major producers of olive with 2,515,800 ha destined to its cultivation (FAOSTAT, 2014). Among the principal cultivars in Spain we can mention 'Arbequina', 'Picual', 'Manzanilla', or 'Cornicabra'. However, other cultivars such as 'Farga' are also important for the production of high quality olive oil and the maintenance of the cultural landscape. 'Farga' is autochthonous from the Maestrazgo region, which constitutes a reserve of millenary trees (Morelló et al., 2004; Belaj et al., 2004a,b).

The present thesis has been performed in the framework of the olive genome project. This project was born as an initiative of three Spanish research institutions: the Centre for Genomic Regulation (CRG), the Royal Botanical Garden (RJB-CSIC) and The National Centre for Sequence Analysis (CNAG); and received the generous support of a private bank (Banco Santander). The main goal of this project was to produce the first reference genome sequence for the Mediterranean olive tree, plus additional sequences from at least one individual per each of the recognized subspecies in the *O. europaea* complex.

## 1.1 Systematics and evolution of the olive tree

Modern taxonomy aims to facilitate the interpretation of the evolutionary history of organisms through an appropriate classification and naming system. Delineating taxon boundaries correctly is crucial because it helps determine whether different individual organisms are members of the same lineage. 'Integrative taxonomy' is defined as the science that aims to delimit the units of life's diversity from multiple and complementary perspectives (phylogeography, comparative morphology, population genetics, genomics, ecology, development, behaviour, etc.). This approach has been used for taxon recognition across the dissertation. In particular, phylogenomic relationships and systematics of the *O. europaea* complex is analysed in detail. In this section I will describe the systematics of *O. europaea* in the context of other plant species and then I will focus on the main studies about the relationships among the different subspecies.

### 1.1.1 Order Lamiales and family Oleaceae

*Olea europaea* L. is an evergreen fruit tree that belongs to the order Lamiales, family Oleaceae, tribe Oleeae, and subtribe Oleinae (Wallander and Albert, 2000; Green, 2004). The order Lamiales is one of the largest orders within angiosperms. It is sister group of the order Boraginales (Refulio-Rodriguez and Olmstead, 2014; Vargas and Zardoya, 2012) and together with other five orders form the clade Lamiids (Figure 1.1). Currently 24 families are recognized within the order Lamiales (Chase et al., 2016), in which the family Oleaceae appears as sister to the family Carlemanniaceae, and is one of the first families that diverged after the family Plocospermataceae (Refulio-Rodriguez and Olmstead, 2014) (Figure 1.2a).

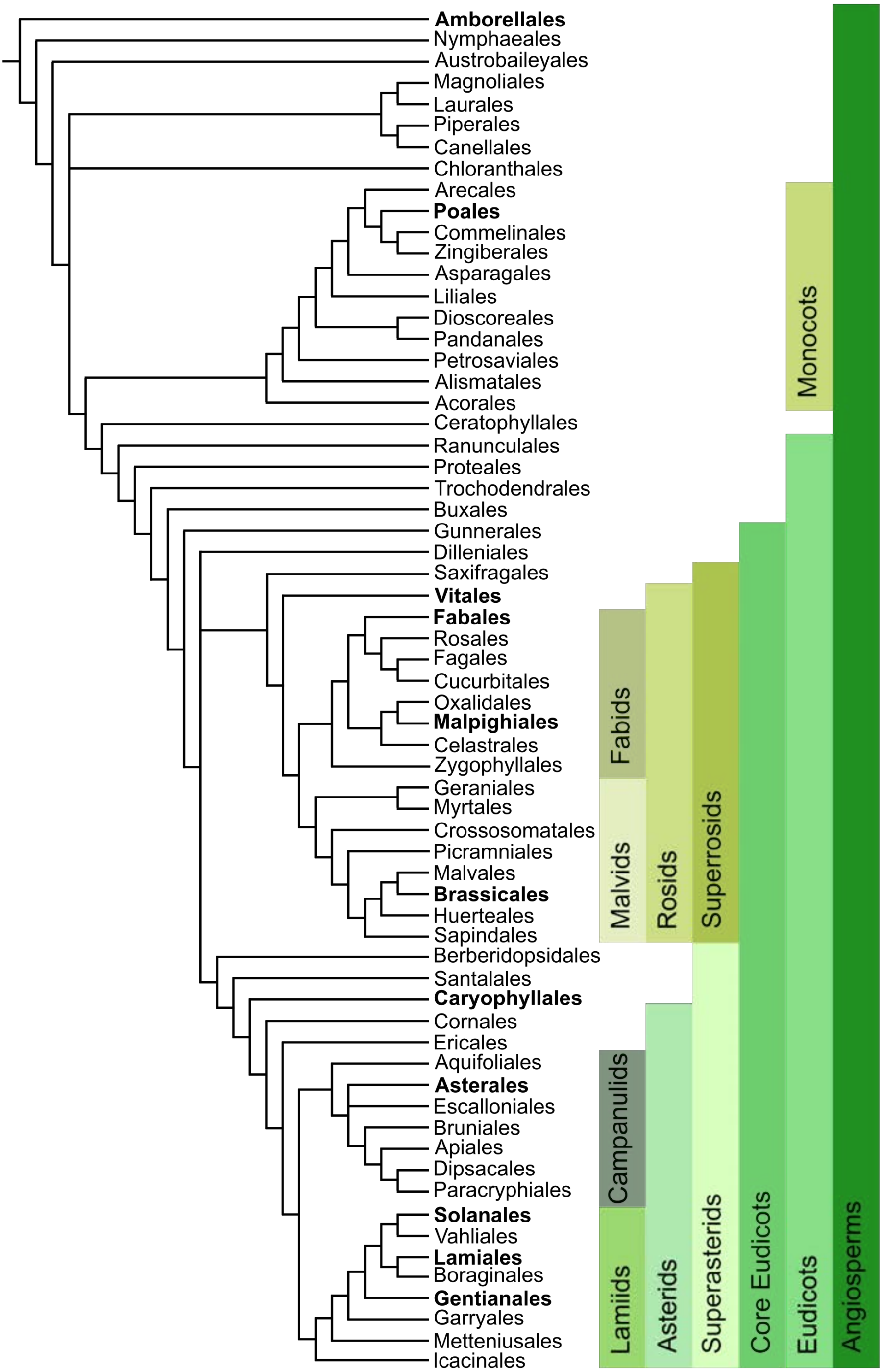

**Figure 1.1:** Phylogenetic tree of the orders of Angiosperms, based on Chase et al. (2016); Refulio-Rodriguez and Olmstead (2014). All the orders marked in bold were included in this study.

The family Oleaceae is composed by five tribes, which vary in their chromosome number from 2n = 22 to 2n = 46 (Taylor, 1945) (Figure 1.2b). The tribe Oleeae is one of the largest groups and its chromosome number is 2n = 46. This tribe is further subdivided into four subtribes: Ligustrinae, Schreberinae, Fraxininae, and Oleinae. This last subtribe comprises thirteen genera, including the genus *Olea*. This genus includes 33 recognized species, of which *O. europaea* is the only cultivated one (Green, 2002, 2004).

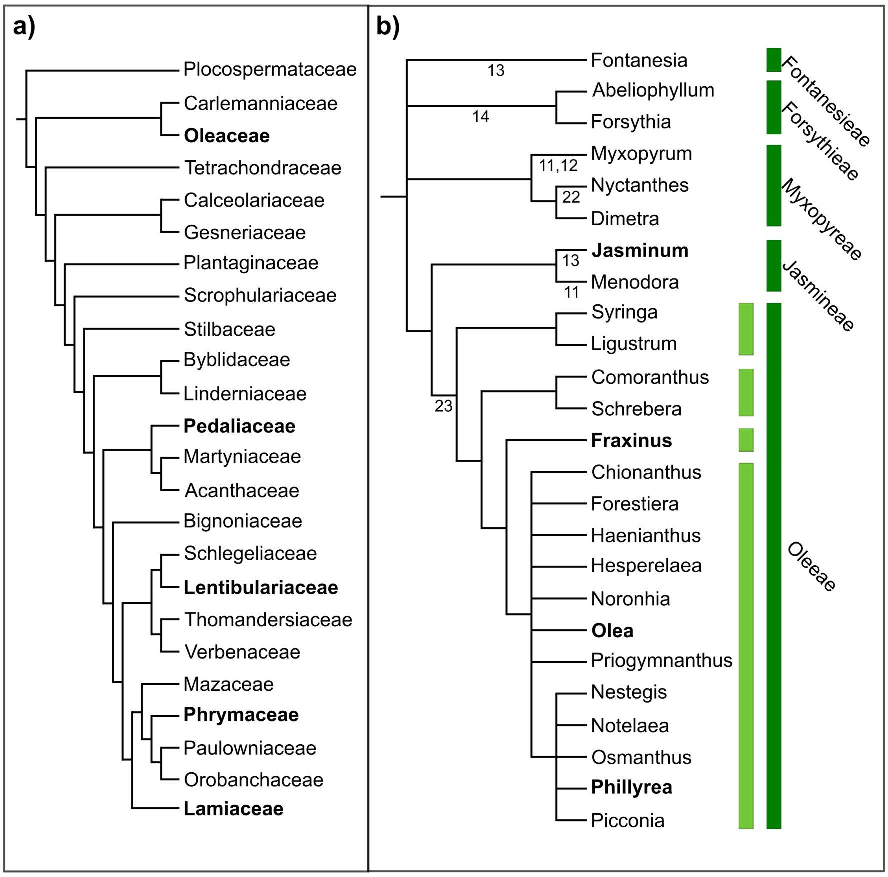

**Figure 1.2:** Phylogenetic tree of the order Lamiales and the family Oleaceae. a) schematic phylogeny of all the families of the order Lamiales (based on Refulio-Rodriguez and Olmstead (2014). b) phylogeny of all the tribes and subtribes described in the family Oleaceae (based on Wallander and Albert (2000)), numbers on the branches indicate the corresponding gametic chromosome number (n). All the clades marked in bold were included in this study.

### 1.1.2 The *Olea europaea* complex

Currently, six subspecies of *O. europaea* are recognised: *europaea*, *maroccana* (Greut. & Burd.) P. Vargas et al., *cerasiformis* G. Kunkel & Sunding, *guanchica* P. Vargas et al., *laperrinei* (Batt. & Trab.) Cif., and *cuspidata* (Wall ex G. Don) Cif. The subsp. *europaea* has two botanical varieties: var. *sylvestris* (Mill.) Lehr. (oleaster), which refers to the wild forms; and var. *europaea* which corresponds to all the cultivated forms (Green, 2002; Vargas et al., 2000). Together, these subspecies constitute the so-called *O. europaea* complex, and each of them show a specific geographical distribution (Figure 1.3): the two varieties of the subsp. *europaea* are distributed in the Mediterranean basin; subsp. *maroccana*, in the Agadir Mountains (Morocco); subsp. *cerasiformis*, in Madeira; subsp. *guanchica*, in the Canary Islands; subsp. *laperrinei*, in Saharan massifs (Hoggar, Ar, Jebel Marra); and subsp. *cuspidata*, from South Africa to southern Egypt and from Arabia to northern India and south-west China (Green, 2002). Currently, *O. europaea* can also be found in Australia, New Zealand, and the Pacific islands because of human-mediated dispersion (Besnard et al., 2007a; Besnard and El Bakkali, 2014).

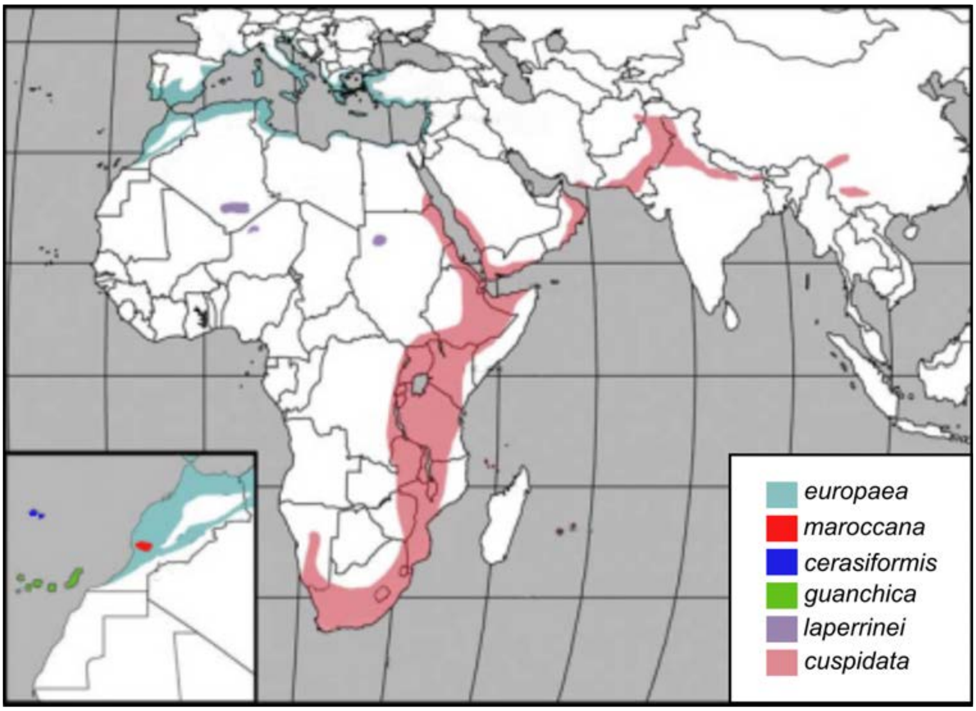

**Figure 1.3:** Geographical distribution of the six *O. europaea* subspecies (taken from Rubio de Casas et al. (2006)). a) Distribution of subsp. *europaea*, *laperrinei* and *cuspidata*. b) Distribution of subsp. *cerasiformis*, *guanchica*, *maroccana* and *europaea* (partial). c) Distribution of var. *sylvestris* in the Iberian Penninsula.

Many cultivated forms of olive trees (cultivars, *O. europaea* var. *europaea*) have been described (Bartolini et al., 1994; Belaj et al., 2004a,b; Trujillo et al., 2014). Olive has preferentially an allogamous mode of sexual reproduction, being most of them self-incompatible, and some male-sterile (Besnard et al., 2000; Breton et al., 2017; Mookerjee et al., 2005). However, because of the long juvenile phase that characterizes the olive tree, these cultivated forms are vegetatively propagated mainly by cutting or grafting (Bracci et al., 2011; Zohary and Spiegel-Roy, 1975). In suitable environments planted clones can persist hundreds or even thousands of years (Rhizopoulou, 2007; Cicatelli et al., 2013)(Figure 1.4). All these characteristics, together with subspecies inter-fertility that results in hybridization, makes the *O. europaea* complex a challenging study group. The correct identification of cultivars and traits of agronomic importance are nevertheless key steps in breeding programs. For instance, an increasing number of studies are focused on the development of molecular markers (Claros et al., 2000; Wiesman et al., 1998; Bandelj et al., 2004; Donini et al., 2006; González-Plaza et al., 2016; Baruca Arbeiter et al., 2014). The development of these markers is also important for food traceability in order to prevent deliberate or accidental mislabeling (Bracci et al., 2011; Raieta et al., 2015).

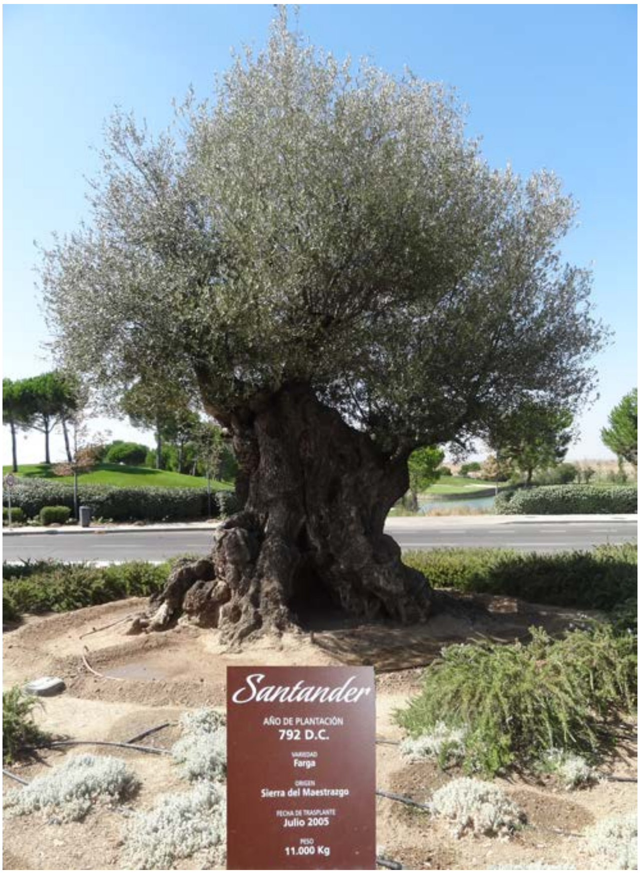

**Figure 1.4:** Cultivated olive tree (*O. europaea* subsp. *europaea* var. *europaea* cv. 'Farga'). This tree, named "Santander", originates from Sierra del Maestrazgo (Spain) and has been estimated to be around 1,200 years old. Leaf material from this individual was used for the sequencing and assembly of the first reference olive genome described in this thesis.

Understanding the taxonomy and diversification of the *O. europaea* complex is important for the management of the genetic resources and for the conservation of genetically differentiated individuals (Besnard et al., 2009). For this purpose the taxonomic limits among the subspecies of the *O.*

*europaea* complex has been long studied using morphological (Green and Wickens, 1989; Médail et al., 2001; Vargas and Kadereit, 2001; Green, 2002) and molecular markers (Besnard and Bervillé, 2002; Besnard et al., 2002a,b; Rubio de Casas et al., 2006; Besnard et al., 2007b, 2009; Diez et al., 2015; Besnard et al., 2011, 2013b).

In plants many phylogenetic and phylogeographic studies are based on organellar genomes because they are abundant, easy to sequence, uniparentally inherited, and they exhibit levels of sequence variation that are reliable for reconstruction of infraspecific relationships (Besnard et al., 2011; Renner and Zhang, 2004; Zhang et al., 2017b; Bernhardt et al., 2017; Li et al., 2016). However, the small amount of polymorphisms in organellar genomes can also be a disadvantage when the purpose is to clarify relationships between closely related lineages (subspecies or even genera) (Diez et al., 2015; Cronn et al., 2002; Small et al., 2004). The nuclear genome is more polymorphic, but the development of nuclear markers can be more challenging (Small et al., 2004). In general the use of both types of markers can help to unravel evolutionary events such as reticulation or incomplete lineage sorting (Besnard et al., 2007b; Linder and Rieseberg, 2004; Petit et al., 2005).

In *O. europaea* both organelles, plastid and mitochondria, are maternally inherited (Besnard et al., 2000). Many studies have been based on organellar markers (Amane et al., 1999; Besnard and Bervillé, 2000; Besnard and Bervillé, 2002; Besnard et al., 2002a,b, 2007b; Besnard, 2008; Mariotti et al., 2010; Besnard et al., 2011; Bronzini de Caraffa et al., 2002) and more recently nine complete plastid genomes have been published for the *O. europaea* complex (Besnard et al., 2011; Mariotti et al., 2010) plus an additional unpublished one has been deposited in the Genbank database (NC _013707). All these studies have allowed the delimitation of seven very differentiated chlorotypes in the *O. europaea* complex, which display specific geographical distributions: E1 (Mediterranean basin and Saharan Mountains), E2 and E3 (the western Mediterranean), M (Macaronesia and southern Morocco), C1 (from eastern Africa to southern Asia), C2 (western Arabia and eastern Africa), and A (tropical and southern Africa) (Besnard et al., 2011, 2007b). Also they show that the most common chlorotype among cultivars was the E1 (Besnard et al., 2011). Similar results were found when using mitochondrial markers (Bronzini de Caraffa et al., 2002; Besnard and Bervillé, 2000; Besnard et al., 2002a).

Many nuclear markers have also been used to assess the genetic variation and the phylogenetic relationships among the different lineages of the *O. europaea* complex (Sebastiani and Busconi, 2017). The principal tools used for these analysis are: ISSR (inter-simple sequence repeats), SSR (simple sequence repeat), RAPD (random amplified polymorphic DNA), amplified fragment length polymorphism (AFLP), ITS (internal transcribed spacer), and SNPs (single nucleotide polymorphisms) (Angiolillo et al., 1999; Hess et al., 2000; Besnard et al., 2001b; Vargas and Kadereit, 2001; Belaj et al., 2003; Besnard et al., 2003, 2007b, 2009; Consolandi et al., 2007; Marchese et al., 2016). Typically ITS has been widely employed in the study of phylogenies in angiosperms (Li et al., 2011) and particularly for resolving phylogenetic relationships in the family Oleaceae (Jeandroz et al., 1997; Wallander, 2008; Li et al., 2002; Besnard et al., 2009). In this context, phylogenetic relationships in the *O. europaea* complex were investigated using ITS (ITS1) of a pseudogene (Besnard et al., 2007b) and functional ribosomal genes (Besnard et al., 2009). The results of these analyses, together with the study of organellar markers, helped to incorporate results of the evolutionary history of olive populations into the current taxonomy of the *O. europaea* complex.

Determining the taxonomic limits among the individuals of the *O. europaea* complex is challenging. One major complication is that genetic barriers do not appear to be significant, neither between cultivated (var. *europaea*) and wild forms (var. *sylvestris*), nor between subspecies (Contento et al., 2002; Besnard et al., 2001b, 2009; Cáceres et al., 2015). Moreover, geographic isolation appears to be the major factor responsible for the observed patterns of differentiation in the *O. europaea* complex (Rubio de Casas et al., 2006). With regards to the phylogeny of this group, it is generally accepted that *O. europaea* is a monophyletic group and the subsp. *cuspitada* is the earliest diverging lineage (Rubio de Casas et al., 2006; Besnard et al., 2009). With respect to the relationships among the other subspecies the picture is less clear, with different results obtained depending on the used marker (Rubio de Casas et al., 2006; Angiolillo et al., 1999; García-Verdugo et al., 2009; Besnard et al., 2002b, 2007b). In this respect, incongruences between phylogenies derived from organellar and nuclear markers have been reported. This incongruences were interpreted as the result of hybridizations between the different subspecies (Rubio de Casas et al., 2006; Besnard et al., 2007b). In this context, in order to better understand the phylogenetic relationships among the individuals of the *O. europaea* complex, we sequenced the whole nuclear genome of at least one individual per each of the described subspecies.

## 1.2 Domestication of the olive tree

The development of agriculture started around 10,000 years ago in close association with the domestication of cereals. Agriculture is considered a milestone in the history of human civilization, and the major cultural development in the last 10,000–13,000 years (Diamond, 2002; Smith, 2011). Domestication is a prerequisite for agriculture (Zeder, 2015), and can be described as a complex process of artificial selection and propagation of an organism to serve as a source of food or other resources of interest (Darwin, 1859; Zohary and Spiegel-Roy, 1975). Cereals were the first plants to be domesticated, and among them one can find the principal 'founder crops' that started food production in south-west Asia and Europe (Old World's civilization). Illustrative examples of such founder crops are wheat and rice (Zohary et al., 2012; Badr et al., 2000; Purugganan and Fuller, 2009). In the Old World, fruits also constituted an important element for food consumption, where olive (*O. europaea* L.), grape (*Vitis vinifera* L.), and figs (*Ficus carica* L.) were the major agricultural products of the Near East and the Mediterranean Basin (Zohary and Spiegel-Roy, 1975; Kaniewski et al., 2012). The domestication process of the olive tree, an emblematic species of the Mediterranean basin, is a matter of current debate. Disentangling the history of domestication of olive is challenging, due to the presence of many factors including vegetative reproduction, gene flow from wild relatives, and human displacements (Diez et al., 2015; Breton et al., 2009). In this introductory section we want to discuss the current knowledge about three main questions: when was the olive tree domesticated?, where did it happen?, and did the domestication of this species occurr once or multiple independent times?

### 1.2.1 History

In order to understand the complex history of domestication of the olive tree, we will shortly discuss the earliest evidence of olive cultivation, found in the Chalcolithic Teleilat Ghassu (3700 to 3500 B.C.), north of the Dead Sea, Jordan Valley. This area was considered located outside or marginal to the natural distribution range of olive trees, and the finding of many carbonized olive stones was interpreted as evidence for cultivation, as such high amounts would be difficult to result from naturally occurring trees (Zohary and Spiegel-Roy, 1975). Based on this and other lines of evidence, about 40 years ago the Levant region was proposed as the place where the olive tree was first cultivated (Zohary and Spiegel-Roy, 1975). From this centre of domestication, it was proposed that the cultivated olive tree gradually diffused from east to west, carried by Phoenicians, Etruscans, Greeks and Romans (Terral et al., 2004).

More recently, several studies have focused on understanding this complex history of domestication and further propagation and diversification of olive cultivars. It is generally accepted that the cultivated olive trees (*O. europaea* var. *europaea*) originate from the wild Mediterranean olive, also called oleaster (*O. europaea* var. *sylvestris*), by artificial selection. Evidence shows that the two varieties are similar in terms of distribution (i.e. sympatric distribution) in the Mediterranean basin, ecological requirements, and morphological characteristics (Besnard et al., 2001a; Besnard and Rubio de Casas, 2016; Zohary and Spiegel-Roy, 1975). It is also assumed that other wild individuals may have contributed to the diversification of the cultivated olive trees (Breton et al., 2006). Finally, regarding the date of domestication, based on both archaeological and genetic studies, it is widely accepted that the olive tree domestication started roughly 6,000 years ago (Margaritis and Jones, 2008; Meadows, 2005; Weiss, 2015; Zohary and Spiegel-Roy, 1975; Kaniewski et al., 2012; Zohary et al., 2012; Besnard et al., 2013b).

The number of domestication events is, however, still debated (Besnard and Rubio de Casas, 2016; Díez and Gaut, 2016). Initially, archaeological and genetic analyses using organellar and nuclear markers lead to the proposition of at least two and up to nine distinct domestication events in different areas of the Mediterranean basin (Besnard and Bervillé, 2000;

Breton et al., 2009; Terral et al., 2004). A more recent research, based on the analysis of complete plastid genomes, proposed that the first domestication took place in the northern Levant, followed by human-mediated dispersal across the Mediterranean basin (Besnard et al., 2013b). This study showed that 90% of the cultivars across the Mediterranean basin share the same chlorotype (E1), which originated in the east Mediterranean basin (Besnard et al., 2013b, 2011). More support to this hypothesis is given by other analyses, in which nuclear markers consistently show that the cultivars were mainly assigned to the eastern oleaster genetic pool (Besnard et al., 2013a). However, new studies argue for a second independent domestication event in the central Mediterranean basin (Diez et al., 2015). Furthermore, a recent archaeological study suggest that the domestication of the olive tree was likely the result of a temporary and regional plantation process (Dighton et al., 2017). In general, the olive domestication origin is much more complex than it was assumed, and the question whether it is a single event followed by secondary events of domestication or there are truly independent domestication events, is still open (Besnard and Rubio de Casas, 2016; Díez and Gaut, 2016).

## 1.3 Hybridization

Hybridization plays an important role in plant evolution (Rieseberg, 1997; Renaut et al., 2014). Typically, hybrids are considered the product of crosses between different species. However, nowadays the word hybrids refers also to the offspring of genetically differentiated populations (Rieseberg, 1997). Hybridization can have many evolutionary effects such as heterosis, transgressive segregation, adaptive introgression, reinforcement, and hybrid speciation (Goulet et al., 2017). Introgression, (or introgressive hybridization) is defined as the transfer of genetic material (usually via hybridization and subsequent backcrossing) between divergent species, lineages or populations (Anderson, 1949). Hybridization with introgression or gene flow affects the genetic and phenotipic composition of populations (Goulet et al., 2017). Excessive gene flow can lead to the extinction of rare taxa through demographic swamping and genetic assimilation (Levin et al., 1996; Todesco et al., 2016). However, introgression can also be positive by introducing new, possible adaptive genetic variation into a population (Goulet et al., 2017). For instance, hybridization is employed in breeding programs of domesticated plants to take advantage of heterosis (hybrid vigor), move desirable variation among lineages, generate novel phenotypes, and increased adaptive potential (Goulet et al., 2017; Rius and Darling, 2014). Heterosis has been extensively studied in rice (*Oryza sativa*) (Langevin et al., 1990; Olguin et al., 2009; Anis et al., 2017), maize (*Zea mays*) (Meyer et al., 2007), and cotton (*Gossypium hirsutum*) (Abro et al., 2009).

Hybrid speciation can occur more commonly via duplication of a hybrid genome (allopolyploidy), but also without a change in ploidy (homoploid hybrid speciation) (Soltis and Soltis, 2009; Rieseberg and Willis, 2007a; Yakimowski and Rieseberg, 2014). Allopolyploidy tend to occur more likely between more-diverged species than homoploid hybrid speciation (Chapman and Burke, 2007; Rieseberg, 1997). In this section we will focus on homoploid hybrid speciation while allopolyploidy will be discussed in the next section (section 1.4).

Homoploid hybrid speciation is considered unusual for a combination of factors such reduced fitness, hybrid sterility, hybrid breakdown, difficulty to acquire reproductive isolation, and difficulty to be identified (Li et al., 1996; Buerkle et al., 2000; Rieseberg and Willis, 2007a; Burton et al., 2013). It has been proposed that reproductive isolation can be achieved by sorting and fixing genetic incompatibilities (Müntzing, 1930), chromosome rearrangements, segregation and recombination (Grant, 1958; Templeton, 1981). But also the possibility that reproductive isolation results from geographical and/or ecological barriers has been argued (Nieto Feliner et al., 2017; Rieseberg and Willis, 2007a).

Hybrids exhibit a large proportion of novel characteristics when compared to their parents. These characteristics might allow hybrids to spread onto new or extreme ecological niches (Rieseberg, 1997). For example, *Pinus densata* is a hybrid that can live in high mountain environments which is inaccessible to both of the parental species (*P. tabuliformis* and *P. yunnanensis*) (Wang and Szmidt, 1990). *Helianthus* hybrids can live in salt marsh habitat (*H. paradoxus*), xeric habitats (*H. deserticola*), and desert sand dunes (*H. anomalus*) (Lexer et al., 2003; Rieseberg et al., 2003, 2007). Indeed hybridization may also serve as stimulus for the evolution of invasiveness (Ellstrand and Schierenbeck, 2006). Some cases in which hybridization preceded the emergence of successful invasive populations are described in *Pyrus calleryana* (Culley and Hardiman, 2009), *Ulmus pumila* (Hirsch et al., 2017), and even *Olea europaea* (Besnard et al., 2014).

### 1.3.1 Hybridization and *O. europaea*

Hybridization has been observed in the *O. europaea* complex. Earlier studies based on AFLPs revealed that the phylogenetic relationships between these taxa are not completely clear and extensive gene flow among these lineages makes the reconstruction of phylogeographic patterns a difficult task (Rubio de Casas et al., 2006). Other studies based on plastid (microsatellites, restriction sites and indels) and nuclear (ITS-1) DNA polymorphisms showed recurrent reticulation events in the *O. europaea* complex (Besnard et al., 2007b). A more recent study based on nuclear microsatellite and plastid DNA showed clear admixture between the subsp. *europaea* and *laperrinei* (Besnard et al., 2013a). Furthermore, hybridization has been put forward as an important factor during the invasion of two subspecies (*europaea* and *cuspidata*) in Australia (Besnard et al., 2007a, 2014).

All the lineages of the *O. europaea* complex seem to be inter-fertile, and their allogamous mode of reproduction might have contributed to the extensive gene flow observed among these lineages (Besnard et al., 2007b; Breton et al., 2006; Besnard et al., 2001b). For instance, it was proposed that geographical isolation rather than genetic barriers appear to be more important for the differentiation of the lineages of the *O. europaea* complex. (Rubio de Casas et al., 2006).

### 1.3.2 Methods to identify hybrids

There are many molecular methods described to identify hybrids and to estimate introgression. In this section, we will describe two widely adopted approaches: molecular phylogenetic and population genetic.

### Phylogenetic approach

Phylogenies can be used to identify allo- and homopolyploid hybridization or introgression (Hobolth et al., 2007; Nieto Feliner et al., 2017). However, for recent hybridization processes, the interpretation of phylogenies can be challenging due to the lack of phylogenetic signal. Moreover, introgression involves reticulation, thereby making the reconstruction of evolutionary histories more difficult (Rieseberg and Wendell, 1993). In this context the phylogenetic relationship of hybrid lineages is better represented through net-like trees, or networks (Willyard et al., 2009). There are two types of phylogenetic networks: explicit and implicit. Explicit networks describe explicit evolutionary reticulation events, e.g. hybridization network. Implicit networks aim to capture incompatibilities in the data, e.g. split network (Yang et al., 2014; Solís-Lemus et al., 2016). Unrooted and rooted phylogenetic networks are currently inferred with different software. SplitsTree 4 (Huson and Bryant, 2006) implements different methods for computing implicit and explicit unrooted phylogenetic networks, such as split decomposition (Bandelt and Dress, 1992), neighbor-net (Bryant and Moulton, 2004), median network (Bandelt et al., 1995), and median-joining (Bandelt et al., 1999). Other software, such as PhyloNet (Than et al., 2008), PADRE (Lott et al., 2009), Perl package Bio:PhyloNetwork (Cardona et al., 2008), Dendroscope 3 (Huson and Scornavacca, 2012) and Julia package PhyloNetworks (Solís-Lemus et al., 2016) computes rooted phylogenetic networks.

Incongruence between organellar and nuclear phylogenies can also be indicative of hybridization (Linder and Rieseberg, 2004). This approach has been used to detect hybridization in the clade *Eupersicaria* (Kim and Donoghue, 2008), the tribe Senecioneae (Pelser et al., 2010), the genus *Pilosella* (Fehrer et al., 2007) and *Olea europaea* (Besnard et al., 2007b), among others. Incongruence among gene trees or between gene trees and species phylogenies can also be indicative of hybridization and introgression events (Rieseberg et al., 1996).

### Population genetic approach

Assessing the genetic variability in populations can lead to the discovery of recent hybridizations by uncovering patterns that are differentially shared across compared populations. In such cases the genomic contribution of each of the parental lineages can be explored for each hybrid (hybrid index or admixture proportion) (Twyford and Ennos, 2012). Many methods have been developed to estimate hybrid index based on morphological and genetic characters (Anderson, 1949; Rieseberg et al., 1999; Anderson and Thompson, 2002; Buerkle, 2005). Among these, we highlight the following software tools: HINDEX (Buerkle, 2005), INTROGRESS (Gompert and Alex Buerkle, 2010), NEWHYBRIDS (Anderson and Thompson, 2002; Anderson, 2008), EIGENSTRAT (Price et al., 2006), FRAPPE (Tang et al., 2005), ADMIXTURE (Alexander et al., 2009; Alexander and Lange, 2011), and STRUCTURE (Pritchard et al., 2000; Porras-Hurtado et al., 2013).

## 1.4 Polyploidization

Polyploidization, or whole genome duplication (WGD), is a key mechanism of genome evolution in eukaryotes and probably also in prokaryotes (Ramsey and Schemske, 1998; Wendel, 2000; Soppa, 2013, 2017). It is defined as the doubling of the complete set of chromosomes of an individual and frequently involves unreduced gametes or interspecific hybridization (Leitch and Leitch, 2008; Ramsey and Schemske, 1998; Van de Peer et al., 2017). It has been described in bacteria (Mendell et al., 2008; Pecoraro et al., 2011; Griese et al., 2011), archaea (Breuert et al., 2006; Soppa, 2011), fungi (Marcet-Houben and Gabaldón, 2015; Wolfe, 2015; Campbell et al., 2016), animals (Luo et al., 2014; Shoemaker et al., 1996; Taylor et al., 2001; Session et al., 2016; Logsdon et al., 2017), and frequently in plants (Ramírez-Madera et al., 2017; Sollars et al., 2016; Vlasova et al., 2016; Yang et al., 2015; Cannon et al., 2006; Schmutz et al., 2010; Gebhardt et al., 2003; Arabidopsis Genome Initiative, 2000; Marcussen et al., 2014; Guyot and Keller, 2004; Paterson et al., 2012; Wang et al., 2016b; El Baidouri et al., 2017; Jarvis et al., 2017; Bomblies and Madlung, 2014).

In plants, polyploidy is one of the major forces of adaptation, speciation, and biodiversification (Ramsey and Schemske, 1998; Soltis et al., 2004; Leitch and Leitch, 2008; Zhan et al., 2016). It is particularly common in angiosperms (flowering plants), where all the species share two ancestral polyploidizations (i.e. paleopolyploidizations), one thought to have occurred in the common ancestor of extant flowering plants, and the other shared with gymnosperms (Jiao et al., 2011; Renny-Byfield and Wendel, 2014; Dodsworth et al., 2016). Also numerous flowering plants —an estimated 25–30% of extant flowering plants— have undergone more recent polyploidizations (i.e. neopolyploidizations) (Soltis, 2005; Bomblies and Madlung, 2014; Ramírez-Madera et al., 2017; Sollars et al., 2016; Jarvis et al., 2017; Van de Peer et al., 2017). Therefore, polyploidy is not a rare event, but it is an ancient and ongoing process contributing to plant evolution (Ramsey and Schemske, 1998; Wendel, 2000; Zhan et al., 2016).

Polyploidy leads to an instantaneous increase in genome size and the complete gene set. In general, gene duplication generates two gene copies and they can be retained (with silencing, sub- and/or neofunctionalization,) or be lost (Leitch and Leitch, 2008; Wendel, 2000; Yu et al., 2017). The most common phenomenon is that duplicated genes accumulate mutations and thereby one of the copies is silenced and eventually pseudogenized (Adams and Wendel, 2005; Lynch and Force, 2000; Sehrish et al., 2014). Alternatively, mutations can lead to functional differentiation of the two duplicates through sub- or neofunctionalization. Both processes can occur at a regulatory level (i.e. paralogous genes change their expression pattern with respect to each other) or at the protein function level (both paralogous proteins diverge from each other in terms of their function). Subfunctionalization generally occurs as a regulatory divergence, where the ancestral gene expression becomes partitioned among the duplicated genes in the relevant tissues and/or stages (Force et al., 1999; Lynch and Force, 2000; Gallagher et al., 2016), while neofunctionalization traditionally implies that the protein encoded by one of the paralogous copies acquires a new beneficial function and the other retains the ancestral function (Lynch and Force, 2000; Walsh, 1995; Gallagher et al., 2016). Both processes have been associated with the origin of the flower and the evolution of the *MADS-box* genes (Dodsworth et al., 2016; Zahn et al., 2006), the appearance of C4 plants

(Monson, 2003; Sage et al., 2012), the evolution of the *KCS* gene family (Guo et al., 2016), and phytochromes (Rensing et al., 2016).

Polyploidization is a driving force for speciation and the emergence of evolutionary novelties (Ramsey and Schemske, 1998; Eric Schranz et al., 2012; Van de Peer et al., 2017; Zhan et al., 2016; Wood et al., 2009). Polyploids often possess novel characteristics that are not present in their diploid progenitors, and it is speculated that they hold a selective advantage through their increased levels of genetic diversity (Ramsey and Schemske, 1998; Dodsworth et al., 2016; Van de Peer et al., 2017). Some of these new characteristics can allow polyploids to adapt and colonize new ecological niches (Ramsey and Schemske, 1998). It is well known that polyploids have higher tolerance for a broader range of ecological and environmental conditions than diploids (Van de Peer et al., 2017). However, polyploidy can also have detrimental effects on fertility and fitness owing to genomic instability, mitotic and meiotic abnormalities, gene expression and epigenetic changes, chromosomal rearrangements and (retro)transposition (Van de Peer et al., 2017; McClintock, 1984). In this context, it has been proposed that (neo)polyploids will often have a higher risk of extinction than do diploids, leading to an evolutionary dead-end (Mayrose et al., 2009, 2015; Arrigo and Barker, 2012). On the other hand, for the survival and long-term success of polyploids, the availability of new ecological niches or an environmental change is necessary, where they may have an advantage over their diploid progenitors (Van de Peer et al., 2017). The last hypothesis has been supported by the observation that many polyploidization events in plants are associated with mass extinction events (Figure 1.5) (for detail check Van de Peer et al. (2017, 2009b); Fawcett et al. (2009); Lohaus and Van de Peer (2016)).

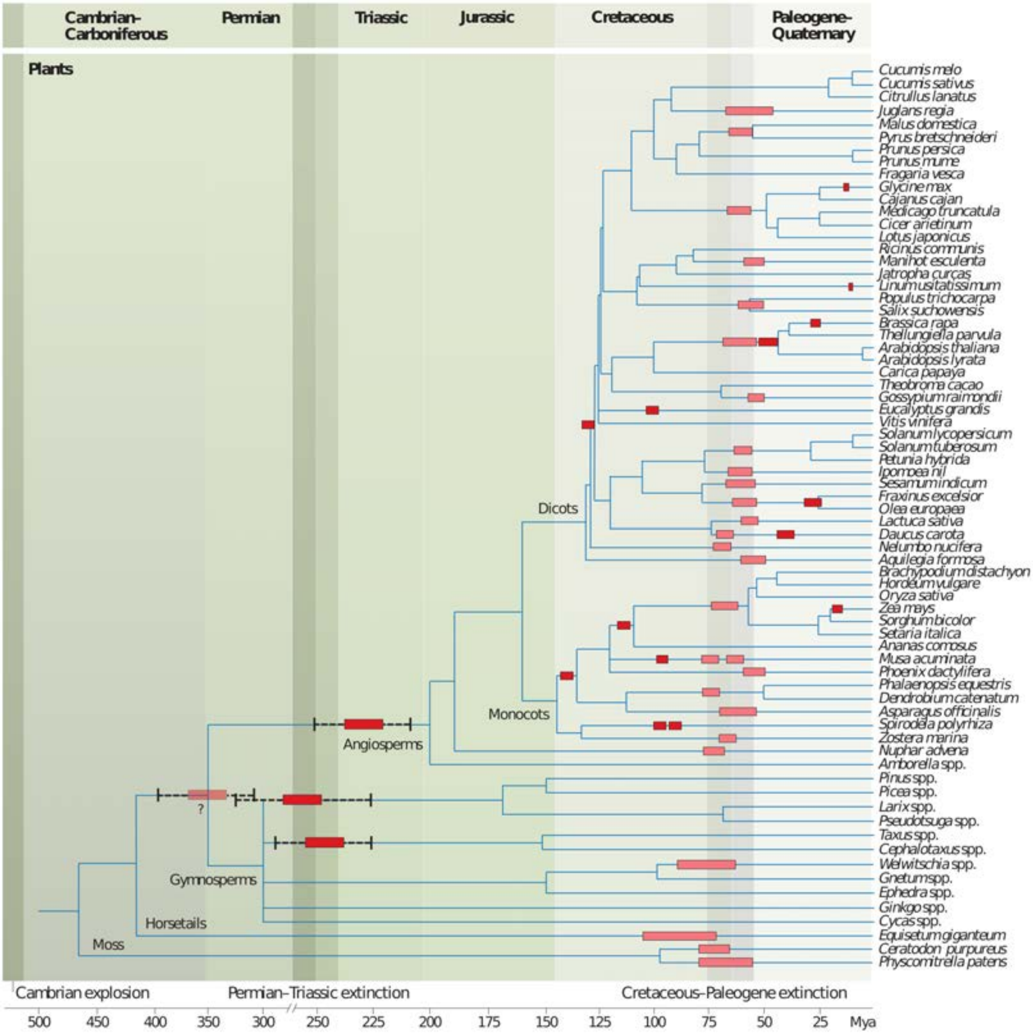

**Figure 1.5:** Phylogenetic tree based on whole-genome duplication (WGDs) data of the plant clade (from algae to angiosperms) (taken from Van de Peer et al. (2017)). Polyploidizations are indicated by rectangles. WGDs estimated to have occurred between 55 and 75 million years ago (Mya) (shaded area around the CretaceousPaleogene boundary) are indicated by light red rectangles. The uncertainty of the date of the events is marked by bold black dashed lines. Mass extinction events are indicated by shaded areas with boundaries 10 million years either side of the predicted date of the event.

### 1.4.1 Autopolyploidization and allopolyploidization

Polyploids can originate by one of two fundamentally different processes: autopolyploidization and allopolyploidization, depending on whether the genome duplicates within the same species or results from the merging two distinct species genomes through hybridization (Ramsey and Schemske, 1998). As a result of these two mechanisms, autopolyploids possess a genome with multiple sets of homologous chromosomes that share a very high similarity, while allopolyploids possess a genome with multiple sets of homoeologous chromosomes, each corresponding to a separate parental genome. A major difference at the cytogenetic level between these two types is the meiotic behavior of chromosomes. Autopolyploids can have multivalent (where more than 2 chromosomes are fully or partially aligned) and random bivalent (pairs) pairings because of the similarity of their homologous chromosomes. Allopolyploids can have mostly bivalent and preferential pairing depending on the divergence between the parental genomes (Ramsey and Schemske, 1998; Chen, 2007; Madlung and Wendel, 2013; Lloyd and Bomblies, 2016). Therefore, autopolyploids exhibits polysomic inheritance, while disomic inheritance is expected to be predominant in allopolyploids (Parisod et al., 2010; Spoelhof et al., 2017). One more characteristic differing the two types of polyploids is the level of morphological differentiation between them and their parentals, and the magnitude of genomic, transcriptomic, and proteomic alterations after polyploidization. In both cases autopolyploids seem to present smaller phenotypic changes, and therefore they are often morphologically similar to their progenitors and with low genomic alterations after polyploidization. In contrast, allopolyploids tend to exhibit intermediate features at both molecular and phenotypic characters as compared to their parents (Barker et al., 2016; Spoelhof et al., 2017).

Multiple pathways (Figure 1.6) can lead to the formation of viable autopolyploids or allopolyploids (Yang et al., 2011; Ramsey and Schemske, 1998). The primary pathway of autopolyploids formation is via the union of unreduced gametes, either through fusion of two unreduced gametes (bilateral polyploidization) or the fusion of reduced and unreduced gametes producing (fertile) triploids that can in turn generate tetraploid progeny through selfing or backcrossing (unilateral polyploidization) (Ramsey and Schemske, 1998; Parisod et al., 2010). Other common pathway in autopolyploids is somatic polyploidization followed by sexual reproduction (Spoelhof et al., 2017). Allopolyploids are presumably formed spontaneously by crossing related species via unreduced gametes or via spontaneous chromosome doubling of the resulting interspecific hybrids (Chen, 2010).

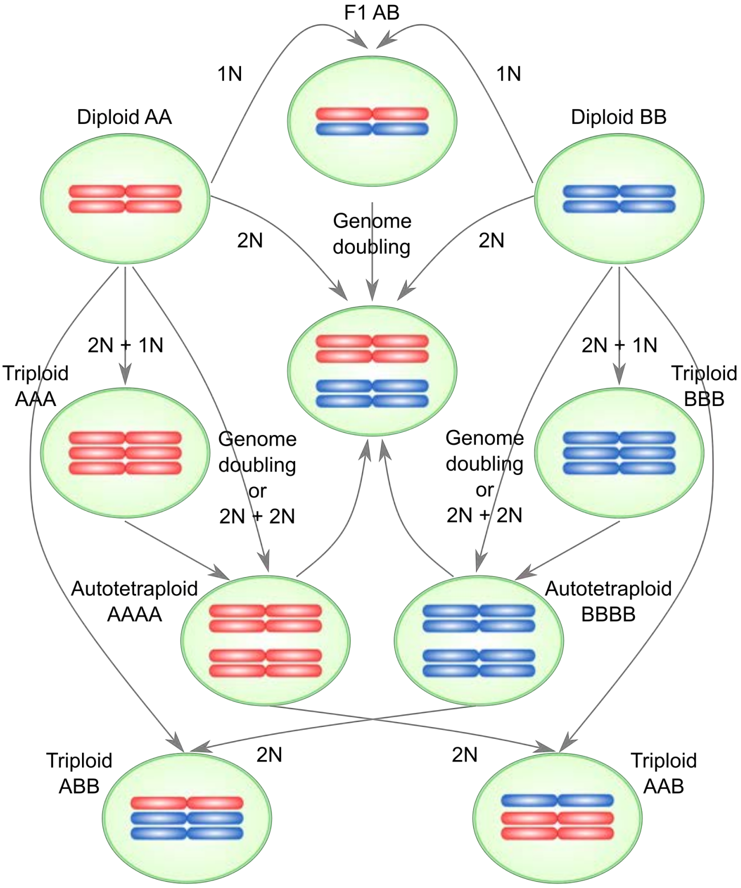

**Figure 1.6:** Main pathways of polyploidization (modified from Yang et al. (2011)).

Both forms, auto- and allopolyploids, are not rare events and have widespread and important evolutionary impacts in flowering plants (Spoelhof et al., 2017; Ramsey and Schemske, 1998; Barker et al., 2016; Parisod et al., 2010; Soltis and Soltis, 2016). Many crop plants are described as allopolyploids: wheat (*Triticum aestivum*) (International Wheat Genome Sequencing Consortium, 2014), cotton (*Gossypium hirsutum* and *G. barbadense*) (Wang et al., 2012; Li et al., 2015a, 2014; Wendel and Cronn, 2001), canola (Brassica napus) (Chalhoub et al., 2014), soybean (*Glycine max*) (Gill et al., 2009; Schmutz et al., 2010), sugarcane (*Saccarum spp.*) (Garsmeur et al., 2011). Although the majority of studies have focused on allopolyploids (Spoelhof et al., 2017), there are also many crops described as autopolyploids. Among the cultivated autopolyploids we can mention sweet potato (*Ipomoea batatas*) (Roullier et al., 2013), potato (*Solanum tuberosum*) (Spooner et al., 2008), alfalfa (*Medicago sativa*) (Havananda et al., 2011), or watermelon (*Citrullus lanatus*) (Saminathan et al., 2015).

### 1.4.2 Polyploidy and domestication

A number of differential traits in polyploids are associated with increment in organ's size (the so-called "gigas" effect), buffering of deleterious mutations, increased allelic diversity and heterozygosity, sub- or neofunctionalization of duplicated genes, and heterosis (hybrid vigor) (Gallagher et al., 2016; Sattler et al., 2016; Renny-Byfield and Wendel, 2014; Hias et al., 2017). The presence of these traits and the ability of polyploids to adapt to new niches may render them more suitable for agriculture than their diploid relatives. In other words, increase of gene diversity of any plant genome offers the opportunity of higher likelihood of survival, morphological diversity and genetic modification. Polyploidy is commonly associated with domestication, and both processes constitute key steps in plant evolution. Recent studies have shown that polyploid species were more likely to be domesticated than their diploid relatives (Salman-Minkov et al., 2016; Fang and Morrell, 2016). Also many crop species are polyploids including banana, canola, potato, wheat, soybean, sugarcane, and cotton (D'Hont et al., 2012; Garsmeur et al., 2011; Li et al., 2015a; Spooner et al., 2008; International Wheat Genome Sequencing Consortium, 2014; Gill et al., 2009; Chalhoub et al., 2014).

Many key phenotypic traits in domesticated plants have a polyploid origin. For instance, in wheat (*Triticum aestivum*) we can find a clear example of the contribution of polyploidization to two important traits for domestication: the grain texture and the free-threshing (Chantret et al., 2005; Wendel and

Cronn, 2001). Domestication of wheat involved three genomes from three divergent species (two genera). In particular, polyploid wheats are the result of two different polyploidization events. In the first one, hybridization between A-genome species (Triticum urartu) and B-genome species (close to *Aegilops speltoides*) gives origin to the tetraploid wheat (*T. turgidum*), and in the second one, hybridization between this tetraploid wheat and a D-genome species (*Aegilops tauschii*) gives origin to the hexaploid wheat (*T. aestivum*) (Matsuoka, 2011). The Hardness (*Ha*) locus represent a clear example of variation after polyploidization. This locus consist of several closely linked genes, and confers the soft grain phenotype in diploid wheat. The deletion of some genes in the tetraploid wheat (*T. turgidum*) resulted in the development of the hard grain phenotype, which is useful for making pasta. After the most recent allopolyploidization, with the incorparation of the D-genome followed by rearrangements in the *Ha* locus, the allohexaploid wheat (*T. aestivum*) resulted in soft grained phenotype (Chantret et al., 2005). The *Q* gene produces a free-threshing character in allohexaploid *T. aestivum* as a result of the functional diversification and interaction between *Q/q* homoeoalleles (from A-, B-, and D-subgenomes) after allopolyploidization. Moreover the free-threshing character is absent among the diploids (Zhang et al., 2011). In cotton, the fibers (single-celled, epidermal, ovular trichomes) of allopolyploids (AD-genomes: *Gossypium hirsutum* and *G. barbadense*) are considerably longer, stronger, and whiter than their diploid relatives (A-genome cottons: *G. herbaceum* and *G. arboreum*, and D-genome cottons: *G. raimondii*) (Wendel and Cronn, 2001; Renny-Byfield and Wendel, 2014). In Brassicaceae, allopolyploids show an increase in the diversity of glucosinolates, because of retention of genes after duplication. These secondary metabolites are powerful weapons in defense against herbivores, a valuable trait in crop species (Hofberger et al., 2013). As compared to diploids, watermelon (*Citrullus lanatus*) autotriploids show more desirable characteristics, such as a higher content of lycopene and citrulline (Liu et al., 2010) and seedless fruits (Chopra and Swaminathan, 1960). In apple (*Malus domestica*) two autotetraploid cultivars perform better in response of two fungal pathogens (Chen et al., 2017). These examples serve to illustrate how polyploidization can provide novel opportunities for selection of key agronomic traits.

### 1.4.3 Diploidization

Despite most angiosperms underwent several ancient or recent WGDs, their genomes are not uniformly large and can vary nearly 2,400-fold (Leitch and Leitch, 2013). This variation results from the fact that the majority of polyploids do not conserve the entire duplicated genome for a long time. Instead, after polyploidization, genome downsizing is the most common response, progressively returning the polyploid genome to a diploid-like state, where chromosomes tend to present diploid behavior with bivalent pairing during meiosis (Wendel, 2000; Leitch and Leitch, 2008; Chen and Ni, 2006). This process is known as diploidization, and involves genomic reorganization: chromosome fusion or loss, (retro)transposon mobility, repetitive DNA loss, and gene loss (Figure 1.7) (Wendel, 2000; Dodsworth et al., 2016; Soltis et al., 2016).

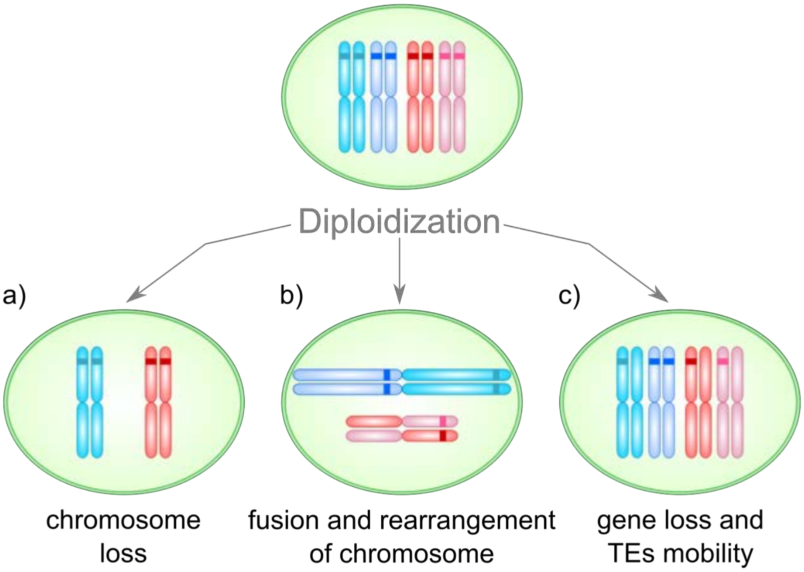
Diploidization
a)
b)
c)
chromosome loss
fusion and rearrangement of chromosome
gene loss and TEs mobility

**Figure 1.7:** Depiction of some mechanism involved in diploidization processes. In this schematic representation we can see chromosome losses (a), chromosome fusions (b), chromosome rearrangements (b), gene loss (c) and retrotransposon mobility (c).

Diploidization may be a requisite for the stabilization of polyploid genomes, and the long-term survival of polyploid lineages. Although it is unknown how fast diploidization can proceed, it is estimated that it can take up to tens of millions of years. This period of time is considered a lag phase required for polyploids to radiate (Dodsworth et al., 2016). Indeed, many studies have demonstrated that plant species with smaller genomes are more likely to have higher diversification rates (Kraaijeveld, 2010). For the genus *Veronica* (Plantaginaceae), it has been proposed that there is a link between genome downsizing and increased diversification following polyploidy (Meudt et al., 2015). Similarly, it has been observed in *Avena* that most of the polyploid species have experienced genome downsizing in relation to their diploid progenitors (Yan et al., 2016). Furthermore, it is commonly observed that species with small number of chromosomes and small genome size may have nevertheless undergone several WGDs in their evolutionary history. For instance, *Utricularia gibba* has a genome size of 82-megabase and underwent three rounds of WGDs (Ibarra-Laclette et al., 2013), *Arabidopsis thaliana* (Brassicaceae) has only 2n = 10 despite having undergone at least two WGDs (Blanc et al., 2003; Bowers et al., 2003). Also gene loss is a common process observed after polyploidization. In *A. thaliana* since the most recent polyploidization event, only 30% of genes retained the duplicate copy (Thomas et al., 2006). In the *Brassica* lineage 35% of the genes inferred to be present when genome triplication occurred have been lost (Town, 2006).

### 1.4.4 Olive tree and polyploidy

In principle, the *Olea europaea* complex involves diploid and polyploid species as revealed by chromosome counting. The number of chromosomes reported for *Olea europaea* is 2n = 46 (Taylor, 1945; Falistocco and Tosti, 1996). Two subspecies are described as polyploids: *cerasiformis* (tetraploid) and *maroccana* (hexaploid) (Besnard et al., 2008), while the other four subspecies (*europaea*, *guanchica*, *laperrinei*, and *cuspidata*) are described as diploids (Baali-Cherif and Besnard, 2005; Besnard et al., 2008). In the subsp. *laperrinei* it was also shown the coexistence of two ploidy types (diploid and triploid genotypes) (Besnard and Baali-Cherif, 2009).

Furthermore, an earlier allopolyploidization event was proposed for the clade of *Olea* and another genera (Taylor, 1945; Falistocco and Tosti, 1996). Recent studies proposed a recent WGD shared between olive and *Fraxinus excelsior*, suggesting an Oleaceae-specific WGD (Sollars et al., 2016). Other studies suggested that a WGD could have taken place at the base of the order Lamiales (Ibarra-Laclette et al., 2013). However these studies did not include the olive genome, and therefore the question of which events really happened in the evolutionary history of olive was still open. Disentangling the complex history of past polyploidization events in the lineage leading to the olive tree has been one of the main foci of this PhD thesis.

## 1.5 Methods to estimate polyploidy

With the advent of widespread genome sequencing, a growing number of plants have been described as recent or ancient polyploids. Indeed, many species traditionally considered as diploids are actually polyploids (Vision et al., 2000; Bowers et al., 2003). The detection of polyploidy is not an easy task. For instance, the signal of WGD can be blurred and difficult to recognize because many genomic rearrangements set in following polyploidy (see diploidization section 1.4.3). Moreover the difficulty increases if the objective is to distinguish between auto- and allopolyploidy, which turns increasingly difficult to discern as more time has passed since the polyploidization event. Also incomplete genomic knowledge about the parents makes it more difficult and sometimes impossible to differentiate. In the following subsections we will summarize the main approaches that are generally used to detect and/or date polyploidy.

### 1.5.1 Chromosome number and nuclear DNA content

In the XX century, inference of the number of current and ancestral chromosome complements were performed based exclusively on chromosome counting. Indeed, the terminology about polyploidization was provided by former cytogeneticists based on chromosome counting in both mitosis and meiosis, followed by inference of ancestral chromosome numbers (see Stace (2010)). Seminal studies proposed the existence of abundant polyploidization processes and extensive variation in ploidy levels in angiosperms based only on chromosome counting (Stebbins and Jr., 1938; Stebbins, 1971).

The somatic chromosome number is represented by "2n" and the gametic chromosome number by "n". On the other hand the base number of a lineage is represented by "x" and represent the ancestral number of gametic chromosomes. In this context, diploids are given as 2n = 2x, triploids as 2n = 3x, tetraploids as 2n = 4x, and so on (Gregory, 2011; Stace, 2010). In more recent times, estimates of the number of WGDs have been based on chromosome number and/or measurement of DNA content per cell (Leitch and Bennett, 2004; Vargas et al., 1999). The logic behind this approach is that polyploidy is the result of multiplications of entire chromosome sets. Initially, it was proposed that genera or families with x = 12 or higher have a polyploid origin (Stebbins, 1971), and species with number of chromosomes larger than n = 9–14 are polyploids (Rieseberg and Willis, 2007b; Goldblatt, 1979). However this parameter by itself can be largely misleading because of the many genome rearrangements that can follow polyploidy (Gregory, 2011).

Several databases are currently available that provide lists of chromosome numbers for diverse plant species, such as the Chromosome Counts Database (CCDB: http://ccdb.tau.ac.il/) (Rice et al., 2015), or the Index to Plant Chromosome Numbers (IPCN: http://www.tropicos.org/Project/IPCN) (Goldblatt and Lowry, 2011) among other databases that are more specific to regions or taxonomic groups (Watanabe, 2002; Bedini et al., 2010; Jara-Seguel and Urrutia, 2011; Hinsley, 2009). With the availability of this information, some probabilistic models have been developed for the analysis of chromosome number changes along a phylogeny, such as chromEvol (Glick and Mayrose, 2014), ChromoSSE (Freyman and Höhna, 2017), and BiChroM (Zenil-Ferguson et al., 2017).

The quantification of DNA is also used to detect polyploidy and the prevailing method is flow cytometry, which is fast and reliable (Doležel et al., 2007; Castelli et al., 2017). Flow cytometry alone or in combination with other methods has been widely used for ploidy estimation in *Medicago sativa* (Brummer et al., 1999; Şakİroğlu and Brummer, 2011), *Crocus sativus* (Brandizzi and Grilli Caiola, 1998), *Crataegus* and *Mespilus* (Talent and

Dickinson, 2005), *Xanthosoma sagittifolium* (Doungous Oumar et al., 2011) and *Arabidopsis lyrata*, (Dart et al., 2004). Flow cytometry, together with nuclear microsatellites, has been used for quantification of ploidy level in *O. europaea* (Besnard and Baali-Cherif, 2009; Besnard et al., 2008). In particular, diploids, triploids, tetraploids and hexaploids have been inferred for the subspecies and some individuals of the *O. europaea* complex.

### 1.5.2 Synteny

Synteny was traditionally used in genetics to indicate the presence of two or more genes (loci) on the same chromosome (McCouch, 2001). In comparative genomics it is used as well in the form of "conserved synteny" and "shared synteny" to refer to the conserved relative order of genes (or genomic regions) in chromosomes of two or more species derived from a common ancestor (Duran et al., 2009; Abrouk et al., 2010). Speciation leads to the appearance of syntenic regions between different species, while segmental duplications or WGD events (polyploidy) give rise to generalized syntenic regions within the same species (Lyons and Freeling, 2008). Although comparative analysis of synteny is a powerful tool to understand the evolutionary history of genes and genomes, it requires high quality genomes, which is a limiting factor. In addition, genome re-organization breaks up synteny and the signal can get blurred over time being difficult or even impossible to recognize. Several methods are currently available to assess synteny such as SynMap (Lyons et al., 2008) from the CoGe platform (https://genomevolution.org/coge/), SimpleSynteny (Veltri et al., 2016), or PLAZA 3.0 (Proost et al., 2015). These methods are able to distinguish between allo- and autopolyploidy in very recent events and often when parent species are known. However, this information is not always available.

### 1.5.3 Phylogenomic methods

In order to better understand phylogenomic methods and thus evaluate polyploidy, we can first classify them into two general categories, namely age distribution methods, and least common ancestor (LCA) methods. Both categories use the relationships of genes between closely related species to identify and place polyploidization events on a species phylogeny (Gregg et al., 2017). They are powerful tools that not only can detect the WGD event, but in many cases they can also tell apart auto- from allopolyploidy. The advantage of these methods is that the information of the parent species does not need to be available, as the detected patterns also emerge when using data from relatives to the parental lineages.

### Least common ancestor (LCA) methods

After a polyploidization events almost all genes and other genomic sequences are initially duplicated (Spoelhof et al., 2017). In allopolyploids, these duplicated genes (paralogs) come from different species, whereby each copy will be sister, in a gene phylogeny, to the respective orthologous gene in the diploid parental species rather than to each other. In autopolyploids each paralogous copy will be sister to each other in a gene phylogeny. A common approach to uncover such relationships is to build gene trees and map duplication events onto the species tree. Two methods can be used for this purpose: (1) detection of the most recent common ancestor of the species involved in the event on the gene tree plus posterior mapping on the species tree (species overlap method) (Huerta-Cepas and Gabaldón, 2011), and (2) gene tree-species tree reconciliation algorithms (Goodman et al., 1979; Doyon et al., 2010; Akerborg et al., 2009; Jacox et al., 2016).

The species overlap method applied on an entire phylome (i.e. the complete collection of evolutionary histories encoded in a given genome) estimates a duplication ratio per branch in the species tree by dividing the number of duplications that map in each node by all the number of gene trees that contain that node (Figure 1.8). This ratio remains close to zero when few duplications are present, but increases when one or multiple WGDs are found (Huerta-Cepas and Gabaldón, 2011).

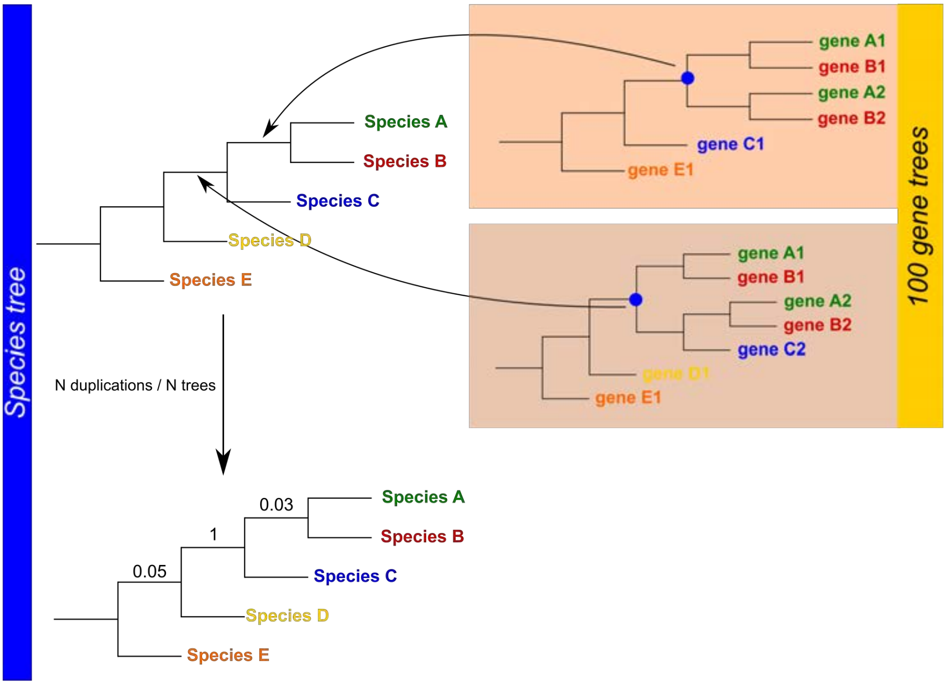

**Figure 1.8:** Schematic representation of the species overlap method (taken from Marcet-Houben and Gabaldón (2015)).

Among the gene tree-species tree reconciliation algorithms we will focus on one recently developed for the analysis of polyploidy: Gene-tree Reconciliation Algorithm with MUL-trees for Polyploid Analysis (GRAMPA). A recent implementation of this software uses the LCA algorithm with multi-labeled (MUL) trees. These MUL-trees are trees in which the tip labels are not necessarily unique, allowing the representation of all sub-genomes in an allopolyploid as descendants of different parental lineages, or as descendants of the same lineage for autopolyploids. This tool allows correct placement of polyploidy events in the phylogeny and distinction between auto- and allopolyploidyzations (Gregg et al., 2017).

In these methods the number of genes that are analysed is important in order to avoid errors produced by an incorrect gene tree inference or incomplete lineage sorting (Gregg et al., 2017). In this context phylomes, that are the complete set of gene trees of a species (Huerta-Cepas et al., 2011), can be used to overcome this problem. For instance, phylome plus the analysis of the ratio of duplications and the analysis of individual gene trees, can be used to distinguish auto- and allopolyploidization. The analysis of phylomes has been used to disentangle the allopolyploid history in *Saccharomyces cerevisiae*, which was believed to be an autopolyploid in the past (Marcet-Houben and Gabaldón, 2015). In this project we used this tool to analyse the polyploidization history in *O. europaea*.

### Age distribution methods

This method is based on the identification of pairs of duplicated genes in the species of interest and the computation, for each duplicate pair, of the number of synonymous substitutions per synonymous site (Ks) (Muse, 1996). These sites evolve in a putatively neutral manner and hence the amount of divergence between any pair of genes is considered to be a good proxy for the age of the duplication (Kimura, 1977; Udall and Wendel, 2006). In the species with duplicated genes and no polyploidization, we will expect that the duplicated pairs are very recent and few will have high Ks, while most of them will show low Ks (Lynch and Conery, 2000). In species with polyploidy, the peaks observed in the distribution generally will correspond to burst of duplications resulting from the WGD events. If the peaks are observed in a single species, it means that the WGD happened in the tip branch of the species tree. Alternatively if peaks of Ks are shared among different species, they generally indicate that the WGD occurred in a common ancestor of the affected species, e.g. (Kang et al., 2014; Blanc et al., 2003; Myburg et al., 2014; Cannon et al., 2015; Barker et al., 2008, 2009). However, there are cases in which such peaks do not represent WGDs, or in which true WGDs cannot be recognized in this manner, due to the stochastic nature of synonymous substitution, which tends to blur the signal, as well as Ks saturation effects, which may lead to artificial peaks (Blanc, 2004; Vanneste et al., 2013). A variation of this method is the transversion rate at fourfold degenerate sites (4DTv). 4DTv measures the fourfold synonymous third-codon transversions between pairs of genes. This is a more conservative estimate of genetic divergence and should be less susceptible to multiple substitution and more commonly occurring synonymous substitutions (transitions) (Comeron, 1995; Muse, 1996; Li, 1993). However as Ks, 4DTv can also be affected by saturation of DNA substitutions (Tang et al., 2008).

### Allele copy number

When polyploid organisms are analysed, it is difficult to deal with heterozygous variations, i.e. those in which more than one alternative nucleotide (allele) is present at the same genomic position. In this case, it is important to measure allele copy number, or the relative representation (e.g. depth of read coverage) of the different variants found at a given genomic position. For example, when calling SNPs in one heterozygous position in a tetraploid, the finding of an A/T polymorphism could represent AATT or ATTT, among other combinations (Dufresne et al., 2014; Clark and Schreier, 2017). Measuring the allele copy number, specifically the alternative allele (also called as B-allele frequency), allows telling apart these possibilities. If we compute a ratio of the alternative allele copy number, we can use it to determine the ploidy level of an organism by plotting the distribution of these ratios. A diploid sample should have one peak around 0.50, a triploid should have two peaks near 0.33 and 0.66, a tetraploid should have three peaks close to 0.25, 0.50 and 0.75, and so on (Figure 1.9) (Zohren et al., 2016).

A general approach is the use of next-generation sequencing (NGS) data, through counting the number of reads that map into a reference and calculating the allele dosage. Some tools can be used for this purpose, with some differences: HANDS2 (Homeolog-Specific Polymorphisms base Assignment using NGS data through Diploid Similarity) (Khan et al., 2016), Control-FREEC (Boeva et al., 2012), nQuire (Weiß et al., 2017), ConPADE (Margarido and Heckerman, 2015), ploidyNGS (Corrêa dos Santos et al., 2017).

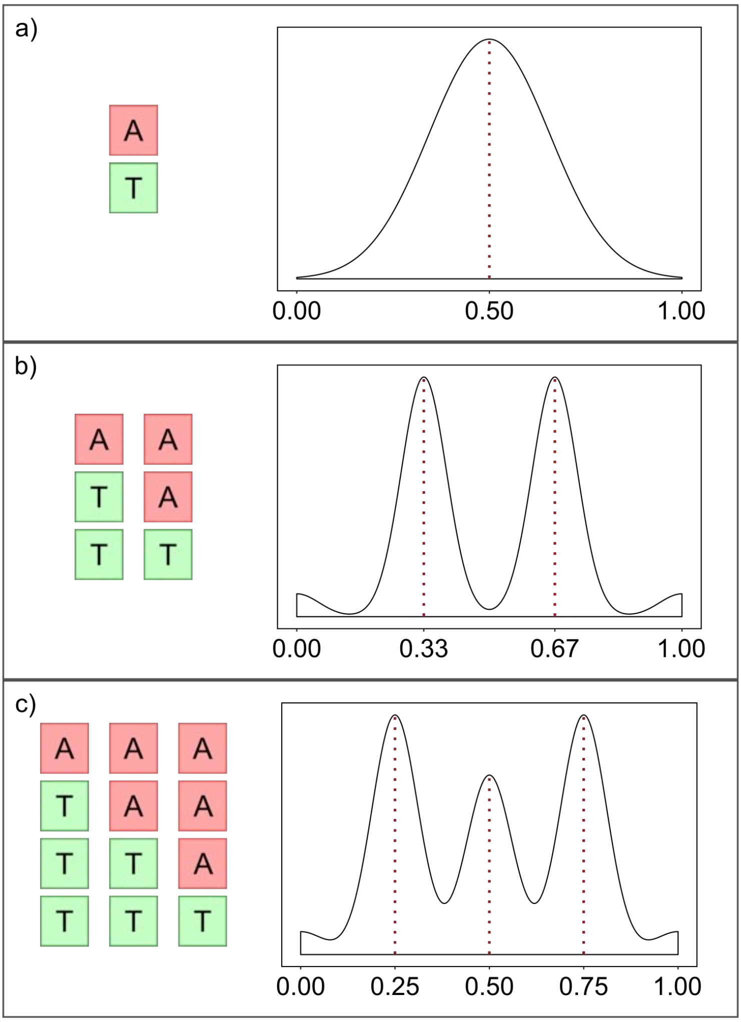

**Figure 1.9:** Density plot for the relative coverage of alternative alleles in heterozygous sites (B-allele frequency) of a diploid (a), triploid (b), and tetraploid (c). Each plot also shows on the left the proportions of heterozygous SNPs per each ploidy.

## 1.6 Final remarks

The olive tree is an emblematic plant of the Mediterranean basin. It constitutes an important source for highly appreciated vegetable oil, an indicator of Mediterranean climate, and a historically-important domesticated species that has accompanied the extension of Mediterranean civilizations. Currently, the olive tree is largely propagated because of the olive oil, and constitutes an economically important crop for many countries of the Mediterranean basin. For this reason many studies have been focused on the application of genetic markers for identification purposes, breeding programs, and the reconstruction of its domestication history. However many questions about its genome evolution, domestication and infraspecific phylogenetic relationships are still debated. The present thesis aims to contribute to a better understanding of these aspects, and to promote future research projects. Our main results are disclosed in **chapters 2**, **3** and **4**. In **chapter 2** we present the first reference genome of the olive tree, together with its annotation and a preliminary comparative analysis with other plant species. In **chapter 3** we used phylogenomics in order to do a deeper comparative analysis of *O. europaea* with other eighteen plant species. Here we unraveled three polyploidization events and we distinguished allo- form autopolyploidizations. In **chapter 4**, we obtained the genome sequences of at least one individual per each of the subspecies of the *O. europaea* complex, and analyse their genetic diversity and phylogentic relationships. In this section another WGD (or partial genome duplication) is unveiled which shortly predate the divergence of the different subspecies of the *O. europaea* complex. Also multi processes of hybridization among these lineages are highlighted in this chapter. In **chapter 5** we offer a summarizing discussion about the main implications of our results and considering future perspectives in these topics. Finally in the **Appendix** section I included the list of publications in which I have contributed.

# Hypothesis

Genomics elucidates plant genome evolution and infers gene duplication by polyploidization and hybridization.

# Objectives

The main objective of this thesis has been to reconstruct and analyze the olive genome, using comparative genomics and phylogenomics methods, in order to increase our understanding of the evolution of olive and pave the way to investigate the genetic bases of traits of agricultural interest.

More specific objectives of my PhD project are:

1. To describe the first reference genome, assemblage and annotation for the olive tree.
2. To investigate the evolution of the olive genome in the context of other sequenced angiosperm species, with a special focus on the characterization of past polyploidization events.
3. To reconstruct the phylogenetic relationships among the infraspecific taxa of the *Olea europaea* complex and estimate evolutionary events within the *Olea europaea* complex.
4. To analyse the genetic history of the domestication itself by comparing genomic variability among cultivated olive tree and its closest wild relatives.

# 2 Olive genome

## Genome sequence of the olive tree, *Olea europaea*

The results of this study have been recently published in the journal *GigaScience*. We herein present the first reference genome of *Olea europaea* L.

Olea europaea is an emblematic species of the Mediterranean basin because it has historically been cultivated for food (olive oil, table olives), a healthy component of the traditional Mediterranean diet. Despite its economic, cultural, and historical importance, the olive tree has been poorly characterized at the genetic and genomic levels. Three Spanish institutions including the Centre for Genomic Regulation (CRG), the Royal Botanical Garden of Madrid (RJB-CSIC), and the National Centre for Genomic Analysis (CNAG) in collaboration, started the olive genome project in 2014. Initially this project aimed to obtain a high resolution sequencing, assembly and annotation of a reference genome. This reference genome will allow an easier and more reliable reconstruction of the evolutionary and domestication history of the olive tree lineages by sequencing additional subspecies and cultivars.

In this chapter we present the work leading towards the reconstruction of the first reference genome of *O. europaea* L. The sequenced individual called "Santander" (cv. 'Farga'), which is a millenary tree, was originally planted in Sierra del Maestrazgo, around the end of the eighth century (Antonio Prieto-Rodríguez personal communication, estimate based on dendrometric analyses). My main contribution to this part of the project was the functional annotation, the analysis of the genomic variation, and the initial comparison with other plant angiosperms.

One of the main achievements is that sequencing of the whole genome will allow genetic improvement and plant manipulation. This will open up numerous research avenues such as olive fruit modification, disease control, breaking of masting, resistence to drought and salinity, among others.

Fernando Cruz*, Irene Julca*, Jèssica Gómez-Garrido, Damian Loska, Marina Marcet-Houben, Emilio Cano, Beatriz Galán, Leonor Frias, Paolo Ribeca, Sophia Derdak, Marta Gut, Manuel Sánchez-Fernández, Jose Luis García, Ivo G. Gut, Pablo Vargas, Tyler S. Alioto and Toni Gabaldón. Genome sequence of the olive tree, *Olea europaea*. *Gigascience*. 2016 Jun 27;5:29. doi:10.1186/s13742-016-0134-5. PMID: 27346392. (*Contributed equally).

# Genome sequence of the olive tree, *Olea europaea*

## 2.1 Abstract

The Mediterranean olive tree (*Olea europaea* subsp. *europaea*) was one of the first trees to be domesticated and is currently of major agricultural importance in the Mediterranean region as the source of olive oil. The molecular bases underlying the phenotypic differences among domesticated cultivars, or between domesticated olive trees and their wild relatives, remain poorly understood. Both wild and cultivated olive trees have 46 chromosomes (2n). A total of 543 Gb of raw DNA sequence from whole genome shotgun sequencing, and a fosmid library containing 155,000 clones from a 1,000+ year-old olive tree (cv. 'Farga') were generated by Illumina sequencing using different combinations of mate-pair and pair-end libraries. Assembly gave a final genome with a scaffold N50 of 443 kb, and a total length of 1.31 Gb, which represents 95% of the estimated genome length (1.38 Gb). In addition, the associated fungus *Aureobasidium pullulans* was partially sequenced. Genome annotation, assisted by RNA sequencing from leaf, root, and fruit tissues at various stages, resulted in 56,349 unique protein coding genes, suggesting recent genomic expansion. Genome completeness, as estimated using the CEGMA pipeline, reached 98.79%. The assembled draft genome of *O. europaea* will provide a valuable resource for the study of the evolution and domestication processes of this important tree, and allow determination of the genetic bases of key phenotypic traits. Moreover, it will enhance breeding programs and the formation of new varieties.

## 2.2 Data description

### 2.2.1 Sequencing

Genomic DNA was extracted from leaf tissue of a single Mediterranean olive tree (*Olea europaea* L. subsp. *europaea* var. *europaea* cv. 'Farga'; NCBI Taxonomy ID: 158383). This tree, named 'Santander', was translocated from the Maestrazgo region (Eastern Spain) to Boadilla del Monte (Madrid, Spain) in 2005. *O. europaea* is a common tree in Spain and there are no legal restrictions on its use for research, including cv. 'Farga'. The tree age was estimated to be 1,200 years old based on dendrometric analyses (Antonio Prieto-Rodríguez personal communication). A combination of fosmid and whole genome shotgun (WGS) libraries were sequenced using Illumina sequencing equipment.

The standard Illumina protocol was followed, with minor modifications to create short-insert paired-end (PE) libraries (Illumina Inc., Cat. # PE-930-1001), which were run on different types of Illumina sequencers (MiSeq 2250, 2300, 2500, 1600 and HiSeq2500 2150) according to standard procedures. The MiSeq XL modes (2500 and 1600) were carried out according to the MiSeq modifications reported in (Birol et al., 2013) and with the technical support of Illumina. Primary data analysis was carried out using the standard Illumina pipeline (HCS 2.0.12.0, RTA 1.17.21.3). Mate-pair (MP) libraries (3, 5, 7 and 10 kb fragment sizes) were constructed at the CRG sequencing unit according to the Nextera Mate Pair Preparation protocol (Illumina Inc.), and sequenced on the HiSeq2500 platform in 2x150bp read length runs. The number of lanes and raw sequenced outputs for each library are summarized in Table 2.1.

**Table 2.1:** Sequencing libraries and respective yields used for whole genome shotgun sequencing and fosmid pools.

| Library | Mode | Name | Yield (Gb) |
|---|---|---|---|
| PE400 | 2*262 | 837G_B | 8.3 |
| PE400 | 2*312 | 837G_B | 68 |
| PE400 | 2*255 | 837G_B | 8.2 |
| PE560 | 2*312 | 846G_D | 33.9 |
| PE560 | 2*151 | 846G_D | 99.2 |
| PE560 | 2*500 | 846G_E_PCR | 14.1 |
| PE560 | 2*151 | 846G_E_PCR | 46.8 |
| PE725 | 2*151 | 837G_E_PCR | 96.3 |
| PE725 | 1*625 | 837G_E_PCR_2 | 15.2 |
| MP3k | 2*151 | T587 | 33.9 |
| MP5k | 2*151 | T586 | 40.3 |
| MP7k | 2*151 | T585 | 37.6 |
| MP10k | 2*151 | T584 | 42.7 |
| FP PE350 | 2*151 | 1FP to 96FP | 11.3* |

*mean yield

Preliminary kmer analysis of PE data (Figure 2.1) indicated a high level of heterozygosity in this sample. To reduce the risk of separately assembling two different haplotypes from the same locus and including them in the final assembly, a fosmid pooling strategy was chosen similar to the one used for the oyster genome project (Zhang et al., 2012a). A fosmid library of 155,000 clones was constructed in the pNGS vector (Lucigen Corp.). Ninety-six pools of $\sim$1,600 clones each were made, and the purified DNA was used to construct short-insert PE libraries using the TruSeqTM DNA Sample Preparation Kit v2 (Illumina Inc.) and the KAPA Library Preparation kit (Kapa Biosystems) according to manufacturer's instructions. The pools were sequenced using TruSeq SBS Kit v3-HS (Illumina Inc.), in PE mode, 2150 bp, in a fraction of a sequencing lane of the HiSeq2000 flowcell v3 (Illumina Inc.) according to standard Illumina operation procedures. The raw sequence yield per pool was 11.3 Gb on average (SD: 2 Gb), corresponding to $\sim$150 depth. In addition a fosmid-end library was created from the same set of clones using the Lucigen pNGS protocol and run in one lane of a HiSeq2000.

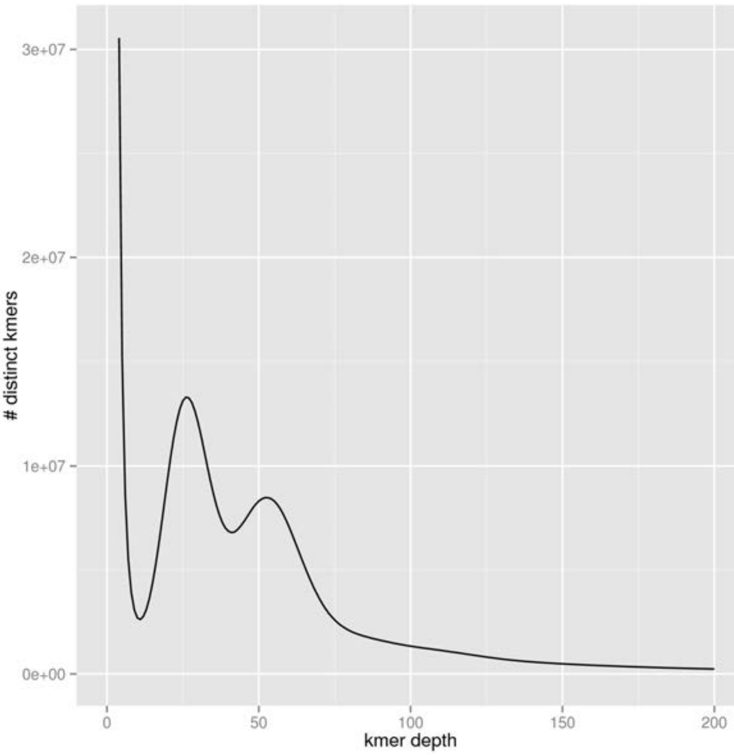

**Figure 2.1:** Kmer spectrum. Using Jellyfish v1.1.10, 17-mers were counted in a subset of whole genome shotgun paired-end reads corresponding to the PE560 2x150 sequencing run. The density plot of the number of unique kmer species (y axis) for each kmer frequency (x axis) is plotted. The homozygous peak is observed at a multiplicity (kmer coverage) of 52 x, while the heterozygous peak is observed at 26 x. The tail extending to the right represents repetitive sequences. The total number of kmers present in this subset was 71,902,584,399. From these data, the Genome Character Estimator (gce) estimates the genome size to be 1.32 Gb.

RNA was prepared from seven different tissues or developmental stages (root, young leaf, mature leaf, flower, flower bud, immature fruit, and green olives), using the Zymo ZR Plant RNA extraction kit (Zymo Research, Irvine, CA). Then, RNA-Seq libraries were prepared using the TruSeqTM RNA Sample Prep Kit v2 (Illumina Inc.) with minor modifications, and libraries were sequenced using the TruSeq SBS Kit v3-HS in PE mode with a read length of 275 bp. Over 50 million PE reads per sample were generated in a fraction of a sequencing lane on a HiSeq2000 (Illumina Inc.), following the manufacturer's protocol. Image analysis, base calling and run quality scoring were processed using the manufacturer's software Real Time Analysis (RTA 1.13.48), followed by generation of FASTQ sequence files using CASAVA software (Illumina Inc.).

### 2.2.2 Genome assembly

A kmer analysis was performed to estimate the genome size, level of heterozygosity and repeat content of the sequenced genome. Using the software Jellyfish v1.1.10 (Marçais and Kingsford, 2011), 17-mers were extracted from the WGS PE reads (PE400), and unique kmers were counted and plotted according to kmer depth (Figure 2.1). The homozygous or main peak is found at a depth of ∼52x. The estimated genome size (found by dividing the total number of kmers by the kmer depth of the main peak) is 1.38 Gb, which is at the low end of the range of empirical estimates. The C-value ranges from 1.452.33 pg (1.42 Gb–2.28 Gb), with the median at 1.59 pg (1.56 Gb) (data from Kew et al. (2012), see Bitonti et al. (1999); Brito et al. (2007); Loureiro et al. (2006, 2007); Ohri et al. (2004)), suggesting the existence of variation in the repetitive fraction of the genome for the species. The left peak at 26x kmer depth indicates many polymorphic sites in the genome. In fact, using the Genomic Character Estimator program, gce v 1.0.0 (Liu et al., 2013), the heterozygous ratio based in kmer individuals is 0.054, and the corrected estimate of genome size is 1.32 Gb. Hereon the gce estimate is referred to as the 'assemblable' portion of the genome.

A pilot WGS assembly using only PE data was performed in order to generate enough contiguous sequences to gather library insert size statistics. PE reads were first filtered for contaminating sequences (phiX, *Escherischia coli* and other vector sequences, as well as *O. europaea* plastids) using GEM (Marco-Sola et al., 2012) with m 0.02 (2% mismatches). Then, the reads were assembled into scaffolds using AbySS v1.3.6 (Simpson et al., 2009) with parameters: –s 600 –S 600-3000 –n 6 –N 10 –k 127 –l 75 –aligner map –q 10. This resulted in an assembly with a total length of 1.94 Gb, and contig and scaffold N50s of 3.7 kb and 3.8 kb, respectively. Library insert sizes were estimated by mapping against this draft assembly. For the WGS PE libraries sequenced on Illumina HiSeq2000 using 2x151 bp reads, the insert size distribution followed a bimodal distribution with a main peak at 725 bp and a smaller peak at 300 bp. Before continuing with the assembly, read pairs belonging to the smaller peak were filtered out, if connecting reads were found overlapping both mates of the pair.

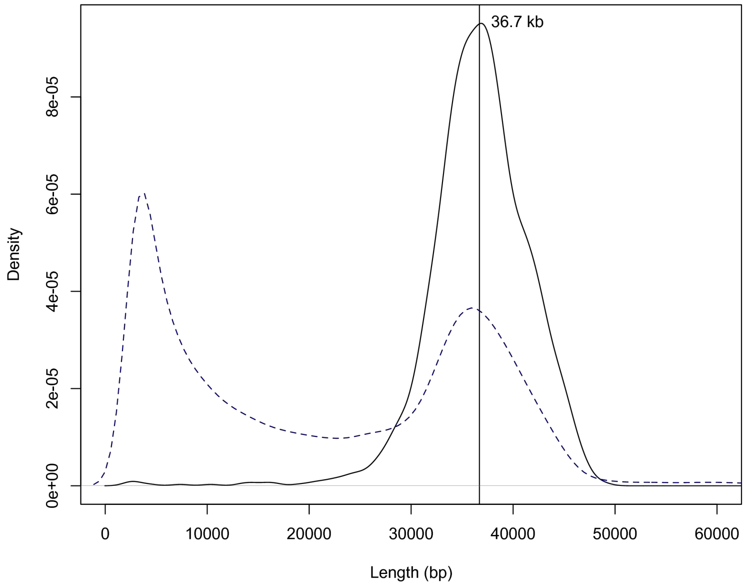

**Figure 2.2:** Comparison of fosmid insert and fosmid-pool scaffold size distributions. Fosmid clone insert size estimates (black contiguous line) were obtained by mapping fosmid end sequences to our merged fosmid pool (FP) assembly. The fosmid end sequencing of only 155,000 unique clones resulted in a very high sequencing depth, so we set a lower threshold of 100 x for the number of times a given length was seen and counted each length only once. While this procedure results in underestimating the amplitude of the density peak, both the shape of the distribution and the mean insert size (36.7 kb) should be unaffected, while the standard deviation is likely an overestimate. The distribution of scaffold lengths from the 96 fosmid pool assemblies is given by the blue dashed line (scaffolds smaller than 2.5 kb were discarded to avoid noise).

The inflated length (47% of the assemblable part of the genome) and the poor contiguity obtained for the draft assembly are symptomatic of the expected difficulty in distinguishing divergent alleles of the same locus from true repeats. To address this challenge, the 96 sequenced fosmid pools (3.9x physical coverage of the genome, each pool covering ∼4% of the genome) were assembled using the assembly pipeline shown in Figure 2.2 to obtain 96 largely haploid assemblies (simulations of 1,600-clone pools with a genome size of 1.38 Gb show a mean of 2.5% of sequenced bases to derive from separate overlapping clones, half of which would come from different alleles). Optimal kmer size was 97 for most of the pools. For each pool a base assembly was produced using ABySSv1.3.7 and parameters: s 300S 3005000n 9N 15k 97l 75aligner mapq 10. Afterwards, the base assemblies went through several rounds of gapfilling (Boetzer and Pirovano, 2012), decontamination, consistency checks, and rescaffolding with ABySSv1.3.7. The decontamination step consists of detecting contaminant sequences (phiX, vectors, UniVec, E. coli, plastids) in the intermediate assemblies using blastn and masking any matches with Ns, thus producing gaps in the assembly. As a result of the FP pipeline, 96 individual assemblies were obtained with an average scaffold N50 of 33,7863,105 bp. The distribution of scaffold sizes follows a bimodal distribution (Figure 2.3), suggesting that a large fraction of fosmid clones are fully assembled. Mapping of fosmid ends to the merged assembly ('FP assembly', see below) gives an estimate of the clone insert size distribution (mean of 36.7 kbSD 4.97 kb) that corresponds well with the right peak of the scaffold sizes.

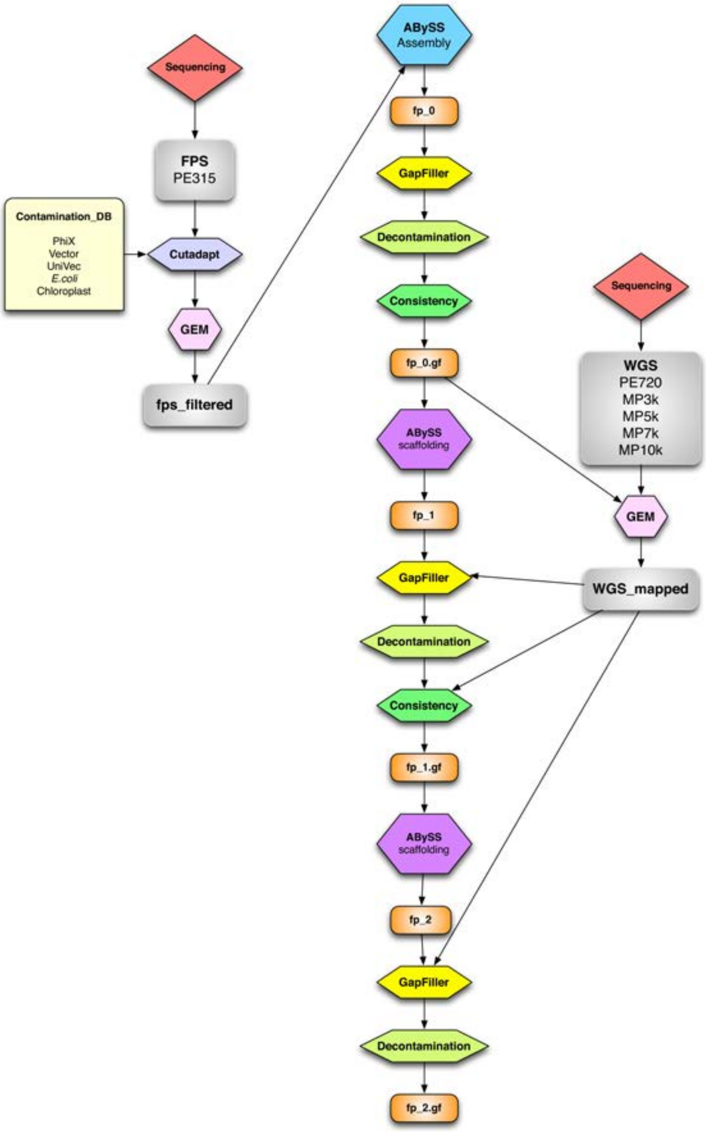

**Figure 2.3:** Fosmid pool assembly pipeline. For each fosmid pool, a single paired-end (PE) library sequenced at 2 x 150 bp was first filtered and trimmed of pNGS vector sequences, as well as those of *Escherichia coli* and other common contaminants, including *Olea europaea* chloroplast sequences. Reads were assembled with ABySS, gapfilled with GapFiller, and contaminants removed using a BLAST homology search. A consistency check was performed, breaking the assemblies at any point inconsistent with the proper insert size and orientation of fosmid pool PE reads. The resulting contigs were scaffolded using whole genome shotgun (WGS) data, followed by another round of gapfilling, decontamination and consistency checking, this time including the new WGS data. To repair the consistency broken assembly, a final round of scaffolding, gapfilling and decontamination was performed.

The 96 fosmid pool assemblies were then merged based on overlaps using in-house OLC-like assembly-merging software called ASM (L. Frias and P. Ribeca, manuscript in preparation; scripts are publicly available at Frias and Ribeca (2016). Two rounds of merging were performed, with intermediate scaffolding and gapfilling steps. In the first round, a minimum overlap of 2,400 bp and high sequence similarity (maximum edit distance of 1.5%) was used, while in the second round, longer overlaps (4,000 bp) and higher sequence divergence (maximum edit distance of 10%) were used in order to merge allelic regions. Each round of merging collapses repeats unless higher order information supports a unique path for resolving a repetitive region; this includes both the sequence of the input data (contigs) and scaffolding information (i.e., the order of contigs in scaffolds in the original fosmid pool assemblies). Merging produced an intermediate assembly (named 'FP assembly' in Figure 2.4) with a scaffold N50 of $\sim$45 kb and a total length of 1.38 Gb. Although this assembly was 4.54% larger than the assemblable genome size (1.32 Gb), gene completeness according to CEGMA was only 95.97% complete and 97.58% partial, suggesting that 2.42-4.03% of the gene space may have been missed.

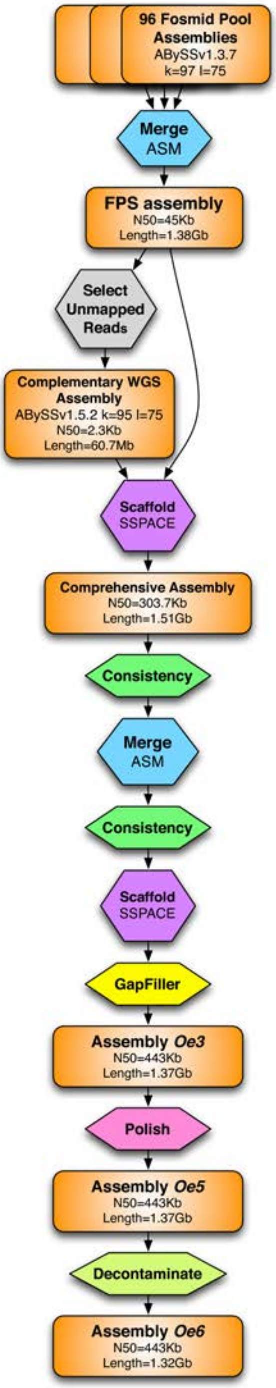

**Figure 2.4:** Overview of the complete assembly pipeline. The basic flow chart starting with the 96 fosmid pool assemblies is shown. Assemblies are shown in orange rounded rectangles. All computational steps are shown as octagons.

To increase the overall completeness of the assembly, all WGS reads that did not map to the FP assembly were selected and used to obtain a complementary assembly using ABySSv1.5.2 with parameters: –s 300 –S 300-5000 –n 10 –N 10 –k 95 –l 75 –aligner map –q 10. This assembly accounts for 60.7 Mbp of sequence, and has an N50 of 1,506 bp for contigs and 2,351 bp for scaffolds. This assembly was then broken into contigs, 50 bp was eroded from the ends of each contig, then contigs smaller than 200 bp were filtered out. Both assemblies were subsequently gathered by joining the WGS contigs with the merged fosmid pool assembly, and scaffolding them with SSPACE 2.0 (Boetzer et al., 2011). To account for read pairs coming from two different alleles in the same genomic region, reads were mapped to the SSPACE input assembly with gem-mapper (settings: m = 0.05 and e = 0.1) and filters were applied to detect unique mappings with no subdominant match. The resulting comprehensive assembly had a scaffold N50 of 303.7 kb and a total length of 1.51 Gb, ∼190 Mb above the expected genome length (1.32 Gb). The excess of assembled sequence is likely to be caused by the presence of artificial duplications during the assembly process (i.e., uncollapsed haplotypes that have been resolved in two different contigs). Several strategies were used to refine the assembly and obtain a haploid reference. First, consistency check was applied to remove local misassemblies by mapping short and intermediate libraries (PE720, MP3k and MP5k) to the input assembly: a positive score is assigned to the assembly regions supported by read pairs separated by distances falling within the limits (mean 3) of the empirical distribution, while a negative score is assigned to regions where read pairs map i) outside of these bounds, ii) in inconsistent orientation, or iii) to different scaffolds. Regions where the sum of these two vectors is negative are removed from the assembly. After applying this consistency check, the resulting assembly had 46,893 consistent contig blocks (compared to 25,042 contigs before the consistency check), giving a total of 1.46 Gb with an N50 of 101 kb. Second, this assembly was collapsed using a minimum overlap of 4 kb and the gem-mapper parameters e 0.03 and m 0.02, so only close matches were merged (similar uncollapsed haplotypes, identical assembly artifacts, and near identical repeats). Additionally, in order to avoid spurious joins, tip merging was applied to the alignment graph down to overlaps of 250 bp.

Finally, no repeat resolution was applied, but coherent links from input scaffolds were reinserted. Consequently, the assembly length shrunk to ∼1.30 Gb, almost matching the assemblable fraction of the genome (1.32 Gb). An additional consistency check was run on the collapsed assembly using the short and intermediate libraries (PE720, MP3k and MP5k), which resulted in breaking the assembly from 64,814 into 72,593 scaffolds, giving a total length of 1.30 Gb with a scaffold N50 of 50 kb. This assembly length is what was expected based on the gce estimate. As a final assembly step, PE reads with high divergence (gem-mapper parameters m = 0.05 and e = 0.08) were mapped to the assembly and rescaffolded with SSPACE 2.0 using parameters k = 3 and a = 0.6. Then, scaffolds shorter than 500 bp were discarded, and the GapFiller program (Boetzer and Pirovano, 2012) was used to close about 40% of the assembly gaps. This assembly was labeled 'Oe3'.

The Oe3 assembly was polished using a mapping-based strategy designed to correct single nucleotide substitution and short insertiondeletion errors. First, one library of paired-end reads (PE725) was aligned using BWA mem (v0.7.7) (Li and Durbin, 2009) and variant calling was performed. Selecting only homozygous alternative variants, an alternative FASTA sequence was obtained using GATK (v3.5) FastaAlternateReferenceMaker (McKenna et al., 2010). After discarding scaffolds shorter than 500 bp, the resulting assembly (Oe5) had a scaffold N50 of 444 kb and a contig N50 of 51 kb. After detecting putative contamination in some scaffolds of the Oe5 assembly, a final decontamination step was performed against yeast, bacteria, arthropod and mitochondrial sequences, combining homology search results obtained by BLAST and, in the case of mitochondrial sequences, regions of high depth (∼6000x). In total, 509 scaffolds were deleted from Oe5 and some parts of another 27 scaffolds were removed. The assembly resulting from this step, Oe6, has a scaffold N50 of 443 kb and a contig N50 of 52 kb (Table 2.2). Oe6 contains 48,419 gaps comprising 53,969,601 sites. The gene completeness of this assembly was estimated using CEGMA (Parra et al., 2007) and BUSCO (Benchmarking Universal Single-Copy Orthologs) (Simão et al., 2015). CEGMA analysis resulted in a gene completeness of 98.79%, while BUSCO, using a plant-specific database of 956 genes, determined a completeness of 95.6% of plant genes. A summary of the complete assembly strategy is shown in Figure 2.4.

**Table 2.2:** Summary statistics of the Oe6 assembly. Numbers of contigs/scaffolds are shown in parentheses.

| Oe6Assembly | Length (bp) | Contiguity (bp) | | | Completeness (CEGMA) | |
|---|---|---|---|---|---|---|
| | | N10 | N50 | N90 | Complete | Partial |
| Contigs | 1,264,682,749 | 138,917 | 52,353 | 11,476 | – | – |
| | (59,457) | (695) | (7,085) | (25,802) | | |
| Scaffolds | 1,318,652,350 | 1,088,680 | 443,100 | 110,965 | 98.8 % | 98.8 % |
| | (11,038) | (94) | (901) | (3099) | | |

### 2.2.3 Partial assembly of an olive tree associated fungus: *Aureobasisium pullulans*

One of the putative sources of non-plant sequence present in the olive samples was considered of interest; it was represented among the fosmid pools and seemed to belong to the fungal genus *Aureobasidium*, which has been previously associated with olive trees (Abdelfattah et al., 2015). To assemble a partial sequence of this genome, four fully sequenced *Aureobasidium* genomes were downloaded from JGI (Gostinčar et al., 2014). Then, BWA v0.7.3a (Li and Durbin, 2009) was used to map all the reads from the fosmid libraries to the four genomes. Once mapped, the reads were filtered allowing only soft clipping for a maximum of one-third of the read, and deleting read pairs when only one of the pairs passed the filters. This resulted in a collection of 18,549,090 reads, which were assembled with SPAdes v.3.1.1 (Bankevich et al., 2012). Scaffolding was done using the assembled fosmids using SSPACE-LongRead (Boetzer and Pirovano, 2014), and gaps were filled with gapcloser (Luo et al., 2012). These two steps were repeated twice. The final alignment was then compared to the *Aureobasidium* genomes using BLAST. Contigs longer than 200 nt, for which less than 20% of their sequence mapped against any of the *Aureobasidium* genomes, were separated and compared against the NCBI non-redundant nucleotide database (Camacho et al., 2009). Only those contigs with first hits to fungal species were kept. The final assembly comprised 18 Mb, roughly two-thirds of the typical size of *Aureobasidium* genomes (2529 Mb). To identify the species and strain, the most common fungal markers used for fungal barcoding were identified (ITS, SSU, LSU, RPB1, RPB2 and EF1). Most of the markers were missing in the assembly or were too short; based on a 769 nt fragment of the RPB1 gene, the most similar sequence was that of *Aureobasidium pullulans* isolate AFTOL-ID 912 (DQ471148.1); a strain that was isolated from the grape plant *Vitis vinifera*. The identity of this fragment was 99.95% indicating that this was likely a different strain of the same species. Augustus (Stanke et al., 2006) was used to perform gene annotation. The training parameters were obtained using scaffold 1 of the published *A. pullulans* genome, and then used to predict proteins in our strain of *A. pullulans*. This resulted in 6,411 proteins.

### 2.2.4 Olive tree genome annotation

To annotate the olive tree genome, consensus gene models were obtained by combining transcript alignments, protein alignments, and gene predictions. A flowchart outlining these steps is shown in Figure 2.5. Transcripts for assembly with Program to Assemble Spliced Alignments (PASA; r2014-04-17) (Haas et al., 2003) were obtained as follows: first, RNA-Seq reads generated from different tissues by our group (see above), plus publicly available datasets in the Sequence Read Archive (SRA) (Table 2.3), were aligned to the final assembly Oe6 with GEM v1.6.1 (Marco-Sola et al., 2012). Transcript models were subsequently generated using the standard Cufflinks v2.1.1 pipeline (Trapnell et al., 2010) starting with the BAM files, resulting in 2,056,606 transcripts, which were then added to the PASA database. In addition, 12,959 olive expressed sequence tags (ESTs) and mRNAs present in Genbank (October 27, 2014) (Galla et al., 2009; Bazakos et al., 2012; Schilirò et al., 2012) were also added to PASA using GMAP v2013-10-28 (Wu and Watanabe, 2005) as the alignment engine. All of the above transcript alignments were then assembled by PASA, resulting in 942,302 PASA assembled transcripts, which were scanned with PASA's Transdecoder program (Haas et al., 2003) to detect likely protein coding regions. This tool predicted a total of 169,562 candidate genes. From these, a training set for ab initio gene predictors was created from PASA models coding for complete proteins, longer than 500 amino acids and with a BLAST hit to either the

Lamiidae or Asteridae proteomes. A training set of 589 non-redundant genes was obtained. In addition, the complete Lamiidae and Asteridae proteomes present in Uniprot (February 10, 2015) were aligned to the olive genome using SPALNv2.1.2 (Iwata and Gotoh, 2012), resulting in 625,980 coding sequence (CDS) alignments.

**Table 2.3:** RNA-Seq samples used for annotation.

| Accession | Tissue | Varietal |
|---|---|---|
| ERS1146989 | Immature olives | Farga |
| ERS1146988 | Roots | Farga |
| ERS1135096 | Old leaves | Farga |
| ERS1135095 | Young leaves | Farga |
| ERS1135094 | Flowers | Farga |
| ERS1135093 | Flower buds | Farga |
| ERS1135092 | Green olives | Farga |
| SRP000653 | Fruits | Coratina |
| SRP005630 | Buds | Picual, Arbequina |
| SRP044780 | Leaves, Roots | Picual |
| SRP016074 | Fruits, leaves, stems and seeds | Picula x Arbequina |
| SRP017846 | Fruits | Istrska belica |
| SRP024265 | Leaves, Roots | Kalamon |

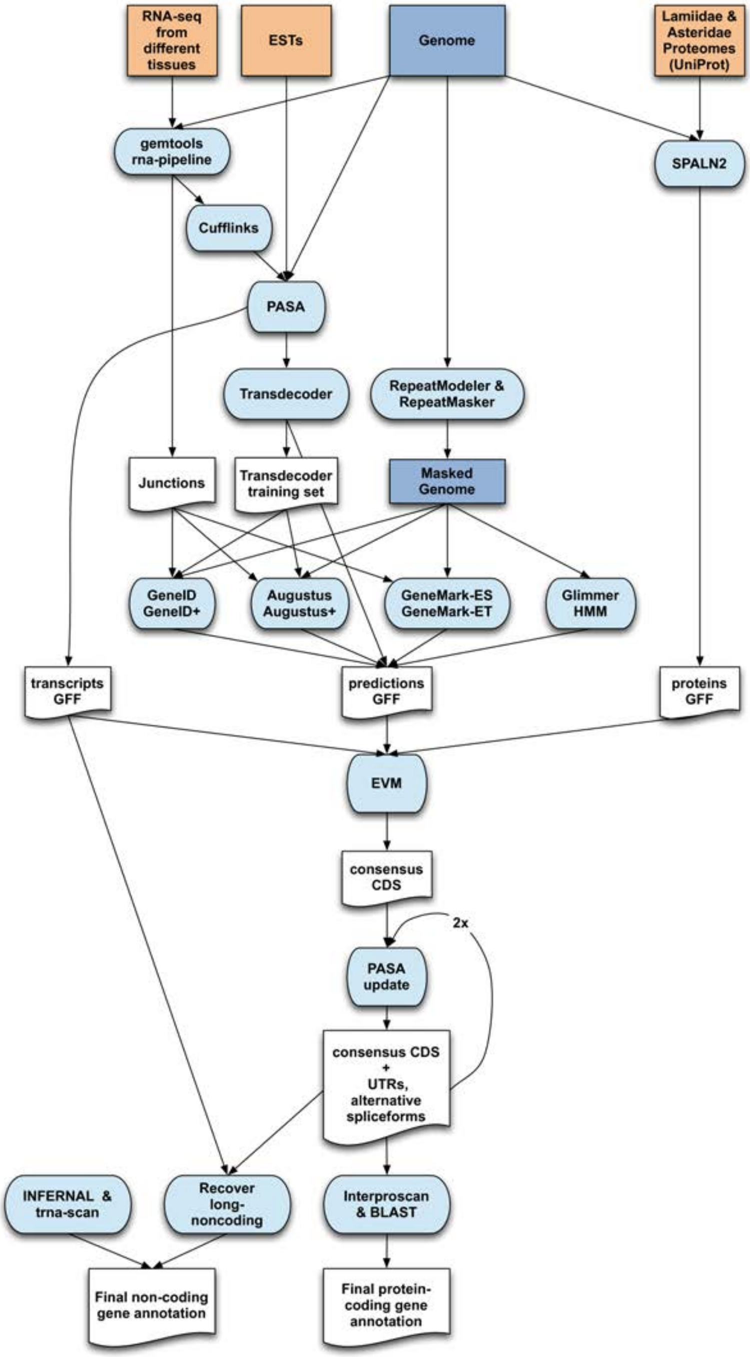

**Figure 2.5:** Overview of the annotation pipeline. Input data for annotation are shown at the top of the flow chart. Computational steps are shown in light blue and intermediate data are shown in white.

For ab initio gene prediction, transposable element repeats in the Oe6 assembly were first masked with RepeatMasker v4-0-5 (Smit et al., 2015) using a custom repeat library constructed by running RepeatModeler v1-0-7 and adding some olive-specific repeats (Barghini et al., 2014). A search was also carried out for masked proteins encoded by transposable elements (TEs) provided in the RepeatMasker Library of TE proteins. Low complexity repeats were left unmasked for this purpose. In total, 63% of the assembly was masked.

On this masked assembly four different ab initio gene predictors were run, since combiners like EvidenceModeler work better when finding consensus among the output of a diverse set of gene prediction algorithms, and orthogonal evidence such as transcript and protein mapping. *O. europaea* protein-coding gene predictions were obtained with GeneID v1.4.4 (Parra et al., 2000) trained specifically for *O. europaea* with GeneidTrainer using the training set of 589 genes; with Augustus v3.0.2 (Stanke et al., 2006) trained with the etraining script that comes with Augustus using the same training set; and with GlimmerHMM v3.0.1 (Majoros et al., 2004) trained with the trainGlimmerHMM script that comes with the program using the same training set. Finally, GeneMark-ES v2.3 (Borodovsky and Lomsadze, 2011) gene predictions were obtained by running it in its self-trained mode. The number of predicted gene models ranged from 48,237 with GeneMark-ES to 97,542 with GlimmerHMM. Geneid, Augustus and Genemark-ET v4.21 were also used to generate predictions incorporating intron evidence, which was extracted from the RNA-Seq data, by obtaining the junctions after mapping it with GEM (see below). Junctions overlapping with ab initio GeneID predictions, Augustus predictions, or with protein mappings were taken as intron evidence. Running GeneID with hints resulted in a total set of 74,231 gene models; Augustus with hints resulted in 70,906; and Genemark-ET with 64,329 gene models.

Evidence Modeler r2012-06-25 (EVM) (Haas et al., 2008) was used to obtain consensus CDS models using the three main sources of evidence described above: gene predictions, aligned transcripts and aligned proteins. EVM was run with three different sets of evidence weights, and the resulting consensus models with the best specificity and sensitivity as determined by intersection (BEDTools v2.16.2 intersect (Quinlan and Hall, 2010)) with the transcript mappings, were chosen for the final annotation (Table 2.4 shows the best-performing weights). Consensus CDS models were then updated with untranslated regions (UTRs) and alternative exons through two rounds of PASA annotation updates. A final quality control was performed to fix reading frames and intron phases, and remove some transcripts predicted to be subject to nonsense-mediated decay. The resulting transcripts were clustered into genes using shared splice sites or substantial sequence overlap as criteria for designation as the same gene. This resulted in a preliminary set of 56,349 protein-coding genes, whose 89,982 transcripts encode 79,910 unique protein products (∼1.59 transcripts per gene). Systematic identifiers with the prefix 'OE6A' were assigned to the genes, transcripts and derived protein products. Functional annotation was performed with InterProScan-5.17-56.0 (Jones et al., 2014), 30,900 protein-coding genes were annotated with gene ontology (GO) terms, and 41,257 were assigned a function.

**Table 2.4:** Weights given to each source of evidence when running Evidence Modeler r2012-06-25.

| Type of evidence | Program | Weight |
|---|---|---|
| ABINITIO_PREDICTION | GeneMark | 1 |
| ABINITIO_PREDICTION | Augustus | 1 |
| ABINITIO_PREDICTION | geneid_v1.4 | 1 |
| ABINITIO_PREDICTION | GlimmerHMM | 1 |
| ABINITIO_PREDICTION | geneid_introns | 2 |
| ABINITIO_PREDICTION | Augustus_introns | 2 |
| ABINITIO_PREDICTION | GeneMark-ET | 2 |
| OTHER_PREDICTION | transdecoder | 2 |
| TRANSCRIPT | PASA | 10 |
| PROTEIN | SPALN | 10 |

The predicted *O. europaea* protein-coding set was then compared with those in four other selected plant genomes (*Arabidopsis thaliana*, *Erythranthe guttata*, *Solanum lycopersicum*, and *Ricinus communis*) downloaded from the NCBI database. A BLASTP search of those proteomes was also performed against the olive proteome, and vice versa, using the BLASTALL 2.2.25+ software suite (Camacho et al., 2009) with an e-value less than 0.01 and with at least 50% of identity (Table 2.5). General statistics for transcript, coding sequence and exon lengths in *O. europaea* are similar to those in the other species, but the number of genes is significantly larger. The number of exons per transcript is slightly lower than in the four compared species. It is possible that more false-positive single-exon genes have been annotated; however, the number of single-exon CDS is not higher, although there is a slight shift in the distribution toward fewer coding exons per transcript (Figure 2.6).

**Table 2.5:** Comparison of *O. europaea* with other plant species.

| Species | Number of proteins | Average transcript length (bp) | Average coding sequence length (bp) | Average exons per transcript | Average exon length (bp) | Proteins with homologs in *O. europaea* | *O. europaea* proteins with homologs in the other species |
|---|---|---|---|---|---|---|---|
| *Olea europaea* | 56,349 | 3,953 | 1,050 | 4.54 | 315 | 56,349 (100 %) | 56,349 (100 %) |
| *Arabidopsis thaliana* | 35,378 | 2,341 | 1,234 | 5.89 | 261 | 23,106 (65.3 %) | 32,796 (58.2 %) |
| *Erythranthe guttata* | 31,861 | 3,378 | 1,351 | 5.77 | 300 | 24,373 (76.5 %) | 42,458 (75.3 %) |
| *Solanum lycopersicum* | 36,148 | 5,626 | 1,389 | 6.48 | 288 | 27,778 (76.8 %) | 38,448 (68.2 %) |
| *Ricinus communis* | 27,998 | 4,323 | 1,390 | 6.53 | 287 | 21,990 (78.5 %) | 37,264 (66.1 %) |

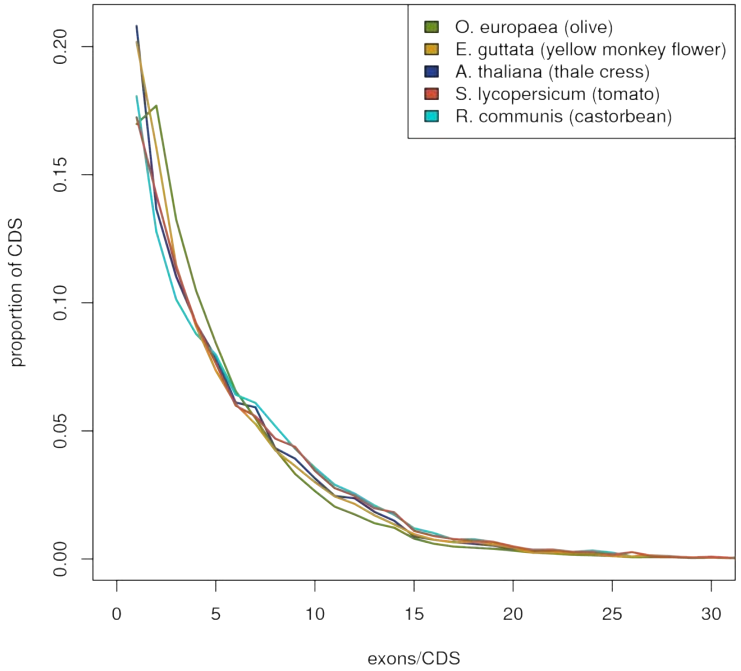

**Figure 2.6:** Distribution of exons per coding sequence in the analysed species. The number of exons per CDS feature (UTRs were ignored) was counted and the distribution plotted for the olive and each of the other four species for which we compared annotations. Similar distributions were observed for all species.

The increased number of coding genes in *O. europaea* suggests the existence of a large-scale genome duplication with respect to the other species. Although this possibility deserves more detailed analysis, preliminary analyses of gene comparisons identified 34,195 *O. europaea* genes with *O. europaea* paralogs that are more similar to each other than to the corresponding best hit in *E. guttata* (80.5% of the total proteins with hits in *E. guttata*), the closest species in this analyses. Also, from the 14,437 paralogous pairs found in *O. europaea* that represent each other's reciprocal best hit, 10,711 pairs had the same best hit in *E. guttata* (which represents 74.2% of the pairs). These results suggest that a high proportion of the *O. europaea* gene repertoire has been duplicated since the separation of these two lamiales species. To discard the possibility that these duplicates resulted from uncollapsed heterozygous alleles, heterozygous single nucleotide variants

(SNVs) identified by variant calling using samtools mpileup in pairs of putatively recent duplicates were counted and compared with those in singletons (genes without recent paralogs). The mean is significantly higher in genes within recent duplicate pairs (Welch's Two Sample t-test p-value <2.2e-16). Finally, the 70% quantile of two-copy SNV counts is 42 and 8 for the one-copy genes. In the case where uncollapsed (duplicated) alleles are frequent, one would expect to obtain the opposite pattern, as reads coming from the same locus would independently map to one of the two uncollapsed haplotypes in the assembly, thus dramatically reducing the number of heterozygous SNVs called. Although further and more detailed analyses are required, these results suggest extensive gene duplication in the lineage leading to the olive tree. The possibility of a whole genome duplication is consistent with the increased chromosomal number in *O. europaea* (2n = 46), as compared to closely related lamiales such as *Erythranthe guttata* (2n = 28) (Fishman et al., 2014) and *Sesamum indicum* (2n = 26) (Zhang et al., 2013).

Non-coding RNAs (ncRNAs) were annotated by running the following steps. First, the program cmsearch (v1.1) that comes with Infernal (Nawrocki and Eddy, 2013) was run with the Rfam database of RNA families (v12.0) (Nawrocki et al., 2015). Also, tRNAscan-SE (v1.23) (Schattner et al., 2005) was run in order to detect the transfer RNA genes present in the genome assembly. To detect long non-coding RNAs (lncRNAs), PASA assemblies that had not been included in the annotation of protein-coding genes (i.e., expressed genes that were not translated to protein) were first selected. Those longer than 200 bp and with a length not covered by a small ncRNA at least 80% were incorporated into the ncRNA annotation as lncRNAs. The resulting transcripts were clustered into genes using shared splice sites or significant sequence overlap as criteria for designation as the same gene. Systematic identifiers with the prefix 'OE6ncA' were assigned to the genes and their derived transcripts. In total, 25,199 non-coding genes have been annotated, among which 20,082 are lncRNAs.

In summary, we report the first genome sequencing, assembly, and annotation of the Mediterranean olive tree. This genome assembly will provide a valuable resource for studying developmental and physiological processes, investigating the past history of domestication, and improving the molecular breeding of this economically important tree.

### Abbreviations

CDS, coding sequence(s); ENA, European Nucleotide Archive; EST, expressed sequence tag; EVM, Evidence Modeler r2012-06-25; FP, fosmid pools; Gb, gigabase; GO, Gene Ontology; lncRNA, long non-coding RNA; MP, mate-pairs; ncRNA, non-coding RNA; PASA, Program to Assemble Spliced Alignment; PE, paired-end; pg, picograms; SNV, single nucleotide variant; SRA, Sequence Read Archive; TE, transposable element; UTR, untranslated region; WGS, Whole Genome Shotgun.

### Acknowledgements

This project was funded by Banco Santander, which provided plant material and financed whole genome sequencing of the olive tree. The authors especially want to thank the late Mr. Emilio Botín for his support in driving this project. IJ was supported by a grant from the Peruvian Ministry of Education: 'Beca Presidente de la República' (2013-III).

### Availability of supporting data

Supporting data are available in the GigaDB database (Cruz et al., 2016b), and the raw data were deposited in the European Nucleotide Archive (ENA) with the project accession PRJEB4992 (ERP004335) for the Olive genome and PRJNA315541 (LVWM00000000) for the *A. pullulans* partial genome. In addition, the genome and annotation can be accessed and browsed at (Cruz et al., 2016c).

# 3 Phylogenomics of the olive tree

# Phylogenomics of the olive tree (*Olea europaea*) disentangles ancient allo- and autopolyploidizations in Lamiales

The results of this chapter has been submitted for publication to BMC Biology, which are now under review. In any case, a preprint has been deposited in the BioRxiv repository and is publicly available (http://www.biorxiv.org/content/early/2017/07/13/163063).

Polyploidy, or whole genome duplication (WGD), is one of the major forces of evolution in flowering plants and a key mechanism for speciation. It is broadly classified in two main types: auto- and allopolyploidy. Autopolyploids initially comprise two nearly identical sets of the same genome, while allopolyploids, result from the merging of two fully differentiated genomes (e.g. different species) followed by a WGD. Understanding the number and type of polyploidization events that a species has experienced until its current condition is not an easy task. In this work we used phylogenomic tools in order to detect, date, and characterize polyploidization events in the course of olive evolution. The olive is an iconic species of the Mediterranean basin and it has been proposed that at least one recent WGD could be involved in its origin. In order to detect and further investigate into this and other putative WGDs, we reconstructed the phylome, i.e. a complete collection of gene evolutionary histories, of the olive and five other Lamiales, which have their genomes already sequenced (*Fraxinus excelsior*, *Mimulus guttatus*, *Sesamum indicum*, *Utricularia gibba* and *Salvia miltiorrhiza*). In the phylome of olive we also included the transcriptomic data of two closely related species (*Jasminum sambac* and *Phillyrea angustifolia*) with the aim of having higher temporal resolution and a more accurate understanding of polyploid events. Our results show that the olive underwent at least three polyploidization events since its divergence from Lamiales: two ancient allopolyploidization events placed at the base of the family Oleaceae and the tribe Oleeae, and a most recent WGD that seems to be specific to the olive lineage. Remarkably, our results show the potential of phylogenomics as an accurate tool to understand the history of polyploidy in plants.

Irene Julca*, Marina Marcet-Houben*, Pablo Vargas, and Toni Gabaldón. Phylogenomics of the olive tree (*Olea europaea*) disentangles ancient allo- and autopolyploidizations in Lamiales. *BMC Biology* (submitted). (*Contributed equally). Available as a preprint in: Irene Julca*, Marina Marcet-Houben*, Pablo Vargas, and Toni Gabaldón. Phylogenomics of the olive tree (*Olea europaea*) disentangles ancient allo-and autopolyploidizations in Lamiales. *bioRxiv*. 2017 Jul 13. (http://www.biorxiv.org/content/early/2017/07/13/163063).

# Phylogenomics of the olive tree (*Olea europaea*) disentangles ancient allo- and autopolyploidizations in Lamiales

## 3.1 Abstract

Polyploidization is one of the major evolutionary processes that shape eukaryotic genomes, being particularly common in plants. Polyploids can arise through direct genome doubling within a species (autopolyploidization) or through the merging of genomes from distinct species after hybridization (allopolyploidization). The relative contribution of either mechanism in plant evolution is debated. Here we used phylogenomics to dissect the tempo and mode of duplications in the genome of the olive tree (*Olea europaea*), one of the first domesticated Mediterranean fruit trees. Our results depict a complex scenario involving at least three past polyploidization events, of which two —at the bases of the family Oleaceae and the tribe Oleeae, respectively— are likely to be the result of ancient allopolyploidization. A more recent polyploidization involves specifically the olive tree and relatives. Our results show the power of phylogenomics to distinguish between allo- and autopolyplodization events and clarify the conundrum of past duplications in the olive tree lineage.

## 3.2 Introduction

The duplication of the entire genetic complement —a process known as polyploidization or whole genome duplication (WGD)— is among the most drastic events that can shape eukaryotic genomes (Vargas and Zardoya, 2014). Polyploidization can be a trigger for speciation (Rieseberg and Willis, 2007a), and can result in major phenotypic changes driving adaptation (Soltis and Soltis, 2016). This phenomenon is particularly relevant in plants, where it is considered a key speciation mechanism (Wood et al., 2009), and where the list of described polyploidizations grows in parallel with the sequencing of new genomes (Fawcett et al., 2009; Xu et al., 2011; Renny-Byfield and Wendel, 2014; Vanneste et al., 2014; Mitsui et al., 2015; Iorizzo et al., 2016). Polyploidization in plants has been a common source of genetic diversity and evolutionary novelty, and is in part responsible for variations in gene content among species (Jiao et al., 2011; Soltis and Soltis, 2016). Importantly, this process seems to have provided plants with traits that make them prone to domestication (Salman-Minkov et al., 2016), and many major crop species, including wheat, maize or potato are polyploids (Xu et al., 2011; International Wheat Genome Sequencing Consortium, 2014; Renny-Byfield and Wendel, 2014)

Polyploidization can take place through two main mechanisms: namely autopolyploidization and allopolyploidization. Autopolyploidization is the doubling of a genome within a species, and thus, resulting polyploids initially carry nearly-identical copies of the same genome. Allopolyploids, also known as polyploid hybrids, result from the fusion of the genomic complements from two different species followed by genome doubling. This genome duplication following hybridization enables proper pairing between homologous chromosomes and restores fertility (Sémon and Wolfe, 2007; Madlung, 2013; Glover et al., 2016). Such mechanism has been described as the fastest (one generation) and most pervasive speciation process in plants (Barker et al., 2016; Doyle and Sherman-Broyles, 2017). Hence, allopolyploids harbor chimeric genomes from the start, with divergences reflecting that of the crossed species.

Elucidating the exact number and type of past polyploidization events from extant genomes is challenging. In part because following polyploidization a process called diploidization sets in, during which the genome progressively returns to a diploid state (Wolfe, 2001). This is attained through massive loss of genes and even of whole chromosomes, resulting in a relatively fast reduction of genome size. For instance, coffee and tomato belong to the class Asteridae. Yet, since their divergence, the tomato lineage underwent a whole genome triplication (Tomato Genome Consortium, 2012). Despite this, the tomato genome encodes only 36% more protein-coding genes than coffee, and has just one additional chromosome. Hence, chromosome number and gene content can serve to point to the existence of past polyploidization events, but are not precise indicators of the number or type of such events. Gene order (also known as synteny) is often used to assess past polyploidizations, generally by comparing the purported polyploid genome to a non-duplicated relative. However, this approach requires well-assembled genomes, and its power is limited for ancient events, as the signal is blurred by the accumulation of genome rearrangements. Finally, phylogenomics provides an alternative approach to study past polyploidizations. In particular topological analysis of phylomes, which are complete collections of gene evolutionary histories, has served to uncover past polyploidization events (Huerta-Cepas et al., 2007; Jiao et al., 2011; Schwartze et al., 2014; Corrochano et al., 2016). Recently, phylome analysis was instrumental to distinguish between ancient auto- and allopolyploidization in yeast (Marcet-Houben and Gabaldón, 2015).

The olive tree (*Olea europaea* L.) is one of the most important fruit trees cultivated in the Mediterranean basin (Besnard et al., 2008). It belongs to the family Oleaceae (order Lamiales), which comprises other flowering plants such as the ash tree (*Fraxinus excelsior*) or jasmine (*Jasminum sambac*). The genome of *O. europaea* has a diploid size of 1.32 Gb distributed in 46 chromosomes (2n). Up to date, polyploids have been described within *O. europaea* as a recent polyploid series (2x, 4x, 6x) based on chromosome counting, flow cytometry and molecular markers (Besnard et al., 2008). However, little is known about ancient polyploidization in the olive tree and relatives. The complete genome sequence has recently been published (Cruz et al., 2016a), with analyses revealing an increased gene content as compared to other Lamiales. This highly suggested the existence of at least one past polyploidization event since the olive tree diverged from other sequenced Lamiales (Cruz et al., 2016a). The recent sequencing of the genome of *F. excelsior* which also presents signs of a past WGD (Sollars et al., 2016), further supports this hypothesis. Still, the exact number and nature of polyploidization events is yet to be resolved. To clarify this puzzle, we performed a phylogenomic analysis of the *O. europaea* genome.

## 3.3 Results and discussion

### 3.3.1 Gene order analysis indicates multiple polyploidizations in the Lamiales

A standard approach to confirm polyploidization relies on the finding of conserved syntenic paralogous blocks. Using COGE tools (Lyons et al., 2008), we searched duplicated genomic regions in the olive genome. Our results revealed numerous such regions, which supports the existence of past polyploidization events (Figure S3.1a). We then calculated the syntenic depth of the olive genome, which is a measure of the number of regions in the genome of interest that are syntenic to a given region in a reference non-duplicated genome (See Methods). As a reference we used *Coffea canephora*. This species belongs to the order Gentianales and, given the presence of duplications among sequenced Lamiales species, *C. canephora* is the closest non-duplicated reference genome (Denoeud et al., 2014). As a control, we performed a similar analysis between *C. canephora* and *Sesamum indicum*, a Lamiales species known to have undergone a single WGD (Wang et al., 2014). We also included *F. excelsior* (Oleaceae) in the comparison as the closest sequenced relative to the olive. Our analyses (Figure S3.1b) revealed contrasting patterns between the three species. The Sesamum-Coffea comparsions showed a clear peak at depth 2, consistent with the reported WGD. In contrast, there was no such clear peak in the above mentioned *Olea*-Coffea or *Fraxinus*-Coffea comparisons, but rather a similarly high number of regions of depth 1 to 6, and 1 to 4, respectively. These results indicate the presence of multiple polyploidization events in the lineages leading to these species, and suggest that *O. europaea* may have undergone more such events than *F. excelsior*.

### 3.3.2 The olive phylome

To elucidate the evolutionary history of *O. europaea* genes and compare it to that of related plants, we reconstructed the phylome (Huerta-Cepas et al., 2011) of this species and those of five other Lamiales (*F. excelsior*, *Mimulus guttatus*, *S. indicum*, *Utricularia gibba* and *Salvia miltiorrhiza*). These phylomes are available in PhylomeDB database (Huerta-Cepas et al., 2014) (see Table S3.1 for details). We reconstructed the evolutionary relationships of the considered species using a concatenated approach with 215 widespread, single-copy orthologs (Figure 3.1a), which yielded congruent results with previous analyses (Wortley et al., 2005; Schäferhoff et al., 2010). We scanned the trees to infer orthologs and paralogs, and date duplication events (see Methods). Using relative dating of gene duplications (Huerta-Cepas et al., 2011) we mapped them to the corresponding clades in the species tree. Functional analyses suggest that phosphatidylinositol activity, recognition of pollen, terpene activity, gibberellin metabolism and stress response are annotations enriched among genes duplicated in several of such periods. We calculated the average duplication frequency for each marked node in Figure 3.1b. Four internal branches showed increased duplication frequencies (nodes 2 to 5). In addition all terminal branches had high duplication frequencies and the two highest frequencies corresponded to the lineages of *U. gibba* (0.53 duplications/gene), for which two recent WGDs have been proposed (Ibarra-Laclette et al., 2013), and to *O. europaea* (0.37). Altogether, these analyses indicate that the lineage leading to the olive tree shows three differentiated waves of massive gene duplications, one preceding the diversification of the sequenced Lamiales (node 4), another at the base of the family Oleaceae (node 5), and another one specific to the olive lineage.

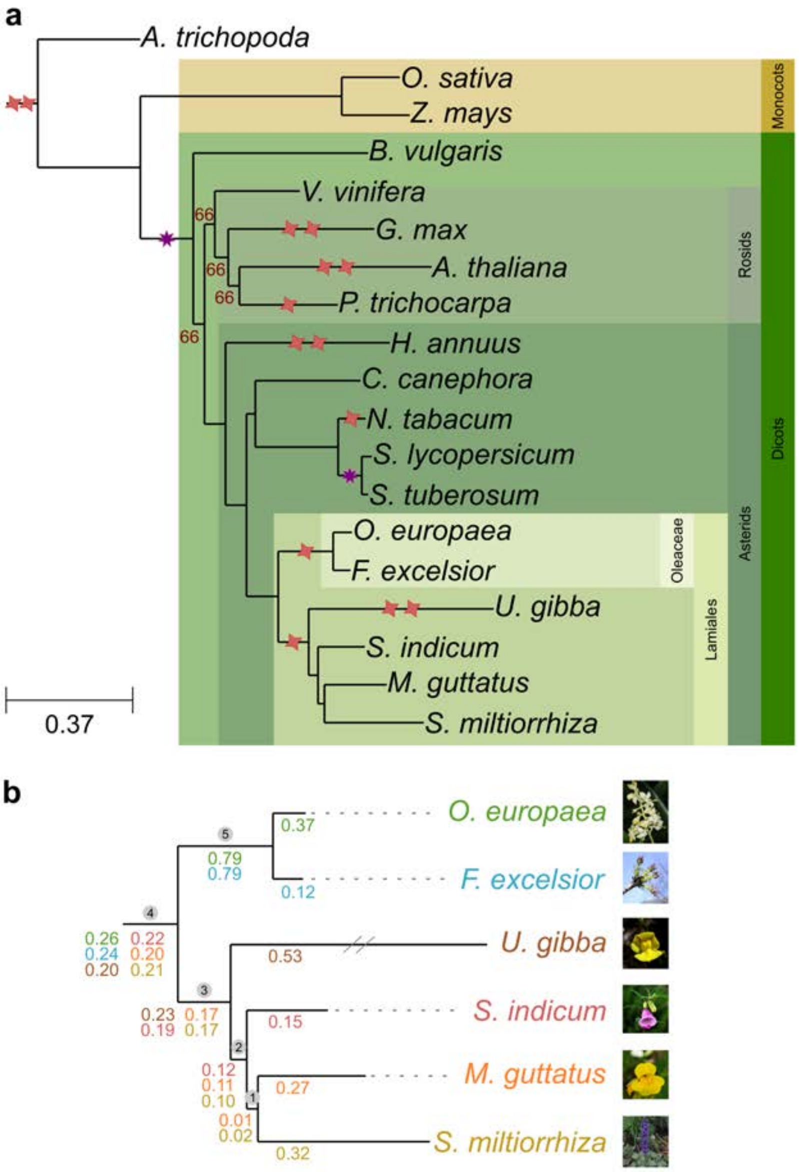

**Figure 3.1:** Species trees. a) Evolutionary relationships between nineteen plants used in this study. All bootstrap values that are not shown in the graph, are maximal (100). Red stars represent WGD events, and purple stars represent whole genome triplication events, as described in the literature. b) Zoom in to the Lamiales clade. Numbers in a circle on top of internal nodes represent the node names as referred to in the text, numbers below each branch are duplication frequencies calculated for each phylome. Each phylome and their corresponding duplication frequencies is colored differently: *O. europaea* - green, *F. excelsior* - light blue, *U. gibba* - brown, *S. indicum* - red, *M. guttatus* - orange, and *S. miltiorrhiza* - yellow.

### 3.3.3 Phylogenetic analysis reveals an ancient allopolyploidization in Lamiales

We focused on the duplication peaks at the internal branches 2, 3 and 4 in Lamiales (Figure 3.1b). A duplication event has been previously described within Lamiales (Hellsten et al., 2013), which could correspond to node 3 or node 4, depending on whether it is shared or not with Oleaceae. The peak at node 2, which has not previously been described, can be explained by the fact that the carnivorous plant *U. gibba*, despite the two recent WGDs, has a reduced genome resulting from massive gene loss (Ibarra-Laclette et al., 2013). Indeed for duplications that occurred at node 3, loss of all the duplicated paralogs in *U. gibba* would lead to mapping to node 2. Supporting such scenario is the finding that, when excluding orphan genes, only 51% of *S. indicum* genes have orthologs in *U. gibba* (see Figure S3.2), as compared to 76% when comparing *S. indicum* to *M. guttatus* (see Figure S3.2). To further test this scenario, we examined trees in the *S. indicum* phylome with node 2 duplications and counted how many of them included *U. gibba* homologs within the Lamiales clade. Only 20.7% of such trees fulfilled that pattern, further supporting that duplications mapped to node 2 mostly result from duplications occurred at node 3 followed by gene loss in *U. gibba*.

A similar scenario could explain duplications at node 3, if massive loss would have occurred in *O. europaea* and *F. excelsior*. Yet, these two species do not have reduced genomes (Figure S3.2). In addition when scanning *S. indicum* phylome trees with either a duplication in node 2 or in node 3, homologs of *O. europaea* or *F. excelsior* could be found in 83.0% of them. Therefore, in this case, losses specific to Oleaceae cannot explain the duplication peak at node 3. This leads to the conclusion that at least two independent duplication events took place in the Lamiales: one corresponds to the previously described event (Ibarra-Laclette et al., 2013; Denoeud et al., 2014) preceding the divergence of *M. guttatus* and *U. gibba* (node 3), and the other, congruent with a more ancestral event (node 4) preceding the divergence between Oleaceae and the other Lamiales species. To further confirm this newly discovered WGD (node 4), we performed a topological analysis on the 10,670 trees in the olive phylome presenting duplications at this node (see Methods), and assessed how many supported each of three possible topologies (see Figure 3.2a): TA.- both paralogous lineages maintain gene copies in at least one species from both Oleaceae and the other Lamiales (non-Oleaceae) species; TB.- One of the paralogous lineages was lost in all non-Oleaceae Lamiales species; and TC.- One paralogous lineage was lost in all Oleaceae species. Surprisingly, many gene trees (77% in the *O. europaea* phylome) supported topology TB (see Figure 3.2b). Equivalent analysis in the other Lamiales phylomes provided consistent results (Figure S3.3).

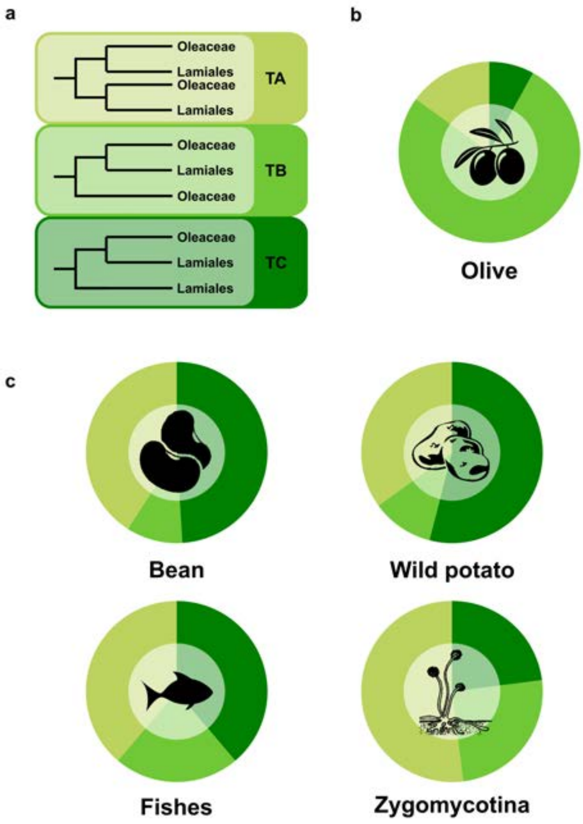

**Figure 3.2:** Topological analysis in olive and four other species. a) Possible alternative topologies after the duplication concerning olive and the other Lamiales. b) Percentage of trees that support each of the topologies shown in Figure 3.2a in the olive phylome. c) Percentage of trees that support each the different topologies for the phylomes of *Phaseolus vulgaris* (bean), *Solanum commersonii* (wild potato), *Scophthalmus maximus* (fishes), and *Rhizopus delemar* (Zygomycotina), taken from PhylomeDB. Like in Figure 3.2a, TB indicates the loss of the paralogous side with the largest amount of species while TC indicates the loss of the paralogous side with the smallest amount of species.

We consider that a preponderance of topology TB is difficult to explain by a simple duplication and loss model. The imbalance in the number of species at the two sides of the node (two Oleaceae vs four non-Oleaceae Lamiales) means that, in scenarios involving gene losses, we expect a greater chance to observe topology TC than topology TB. This expected preponderance of TC was supported in analysis of other phylomes comprising WGD events at a node sub-tending imbalanced clades (see Figure 3.2c). An alternative explanation for the preponderance of TB topology is the presence of an allopolyploidization at the base of Oleaceae. Indeed hybridization between an ancestor from a lineage that diverged before the Lamiales species included in our set and another species more closely related to the non-Oleaceae Lamiales would explain our observation (Wolfe, 2001; Marcet-Houben and Gabaldón, 2015) (see Figure S3.4 for a detailed scenario).

### 3.3.4 Increased phylogenetic resolution provided by transcriptomes uncovers allopolyploidization at the base of the tribe Oleeae

The ability to discern relative timing and type of past polyploidizations depends on the taxonomic sampling of the compared genomes. Unfortunately, at the time of starting this analysis the olive tree and *F. excelsior* were the only fully sequenced genomes from within the family Oleaceae. To increase the resolution of our analyses we included the transcriptomes of different Oleaceae species, whose genomes are not available: *Jasminum sambac* (Li et al., 2015b) and *Phillyrea angustifolia* (Sarah et al., 2017). The two species plus *F. excelsior* represent three important divergence points in the olive lineage. *P. angustifolia* belongs to the same subtribe (Oleinae), *F. excelsior* belongs to the same tribe (Oleeae) and *J. sambac* belongs to the same family (Oleaceae). In addition *J. sambac* has only 26 (2n) chromosomes, whereas the other three species have 46 chromosomes, which suggests that *J. sambac* likely experienced a lower number of polyploidizations. We thus expanded the olive phylome with these transcriptiomes (see Methods). We then selected two sets of trees: namely those including at least one sequence of each newly included species (set1: 20,705 trees) and those where a monophyletic clade contained the olive protein used as a seed in the phylogenetic reconstruction, and at least one sequence of each of the newly included species (set2: 11,352). Using the same approach described above we reconstructed the phylogeny of the expanded set of species (Figure 3.3a), which was congruent with previous analyses based on plastid DNA (Wallander and Albert, 2000). Additionally we estimated their divergence times (see Methods and Figure S3.5). The nodes (branches) in the new phylogeny were named from A to E (Figure 3.3a), where E matched node 4 in the initial species tree (Figure 3.1b). A new duplication profiling using set1 suggests three main duplication peaks in Oleaceae at nodes A, C, and D (see Figure S3.6). The node at the base of the family Oleaceae (node D) is of similar density as the peak found at the base of the Lamiales (node E), which we already described as an allopolyploidization event that happened at the base of the Oleaceae family. Another peak at the base of the Oleeae tribe (node C) is higher than the previous two peaks, as could be expected of a more recent event. A third peak (node A) was still found specifically in *O. europaea*, indicating this duplication occurred after the divergence with *P. angustifolia*. Moreover, when duplication ratios are based on the more stringent set2 (see Figure S3.6), ratios in nodes C and D are affected, while the rest remain with a similar density as in set1. The increased presence of proteins of *J. sambac* sister to the olive protein in gene trees in set2, can explain the increase in the ratio in node D, but not the decrease of the ratio in node C.

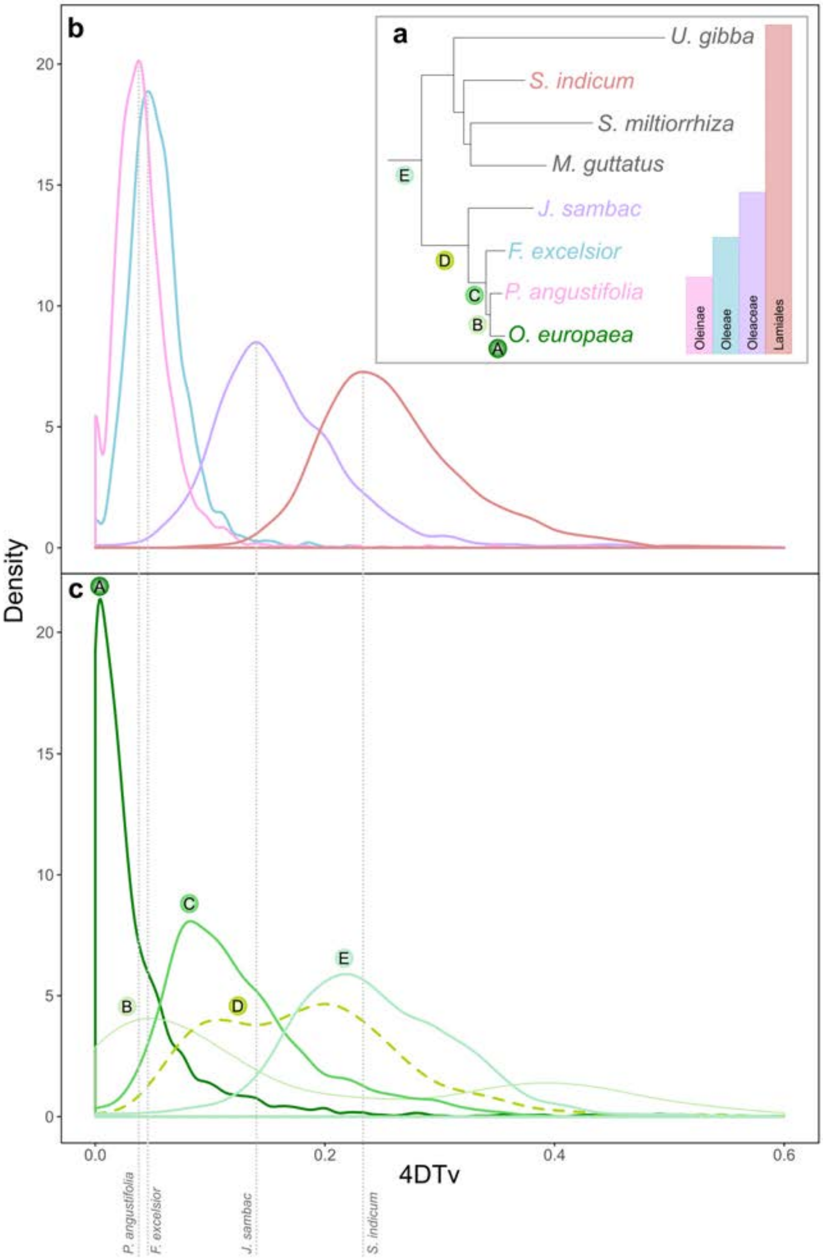

**Figure 3.3:** Species tree and 4DTv of the set1. a) Species tree of the group of Lamiales including the additional two Oleaceae species, bars on the right show the taxonomic classification. Nodes where the 4DTv of the paralogous pairs were calculated are marked with letters (A to E) as referred to in the text and colored according each evolutionary age. The species used to calculate the 4DTv of orthologs pairs are shown in different colors. b) 4DTv of the orthologous pairs between *O. europaea* with *P. angustifolia*, *F. excelsior*, *J. sambac* and *S. indicum*. c) 4DTv of the paralogous pairs of *O. europaea* at the marked nodes in the tree.

To obtain an independent assessment of the relative age of duplications, we plotted the ratio of transversions at fourfold degenerate sites (4DTv) for pairs of paralogs mapped at each of the branches in Figure 3.3a, and compared these ratios with those of orthologous pairs found between *O. europaea* and the three other Oleaceae species plus *S. indicum* (see Figure 3.3 and Figure S3.7). The resulting patterns (Figure 3.3) indicated overall congruence between topological dating and sequence divergence. The youngest peak comprised olive-specific duplications and followed the separation of olive and *P. angustifolia* of ∼10 Mya (see Figure S3.5). The second wave of duplications appeared after the divergence of *J. sambac* and before the divergence of *F. excelsior*, at the base of the Oleeae tribe, which diverged between 14–33 Mya. Interestingly, duplications whose topology maps to two nodes appeared in this region of the 4DTv: those that map at node C after the divergence of *J. sambac* and a fraction of the duplications that happened before the divergence of *J. sambac* (node D). The most ancient duplication wave corresponds to the allopolyploidization event that we have previously described occurred between 33–72 Mya at the base of the Oleaceae family (node D). Of note this time frame includes the Cretaceous-Tertiary (KT) mass extinction event, around which many other plant polyploidization events have been predicted (Fawcett et al., 2009). The fact that duplicatons whose topology map at node E are found in this region of the 4DTv, placed after the divergence of *S. indicum*, further supports the hybridization claim we propose. We also note that part of the duplications mapping at node D are found in this region. Altogether, these results confirm the presence of duplication three waves of duplications but also show that the duplications mapping at node D are divided in two peaks of sequence divergence as indicated by 4DTv plots. Node D duplications with 4DTv values found between the divergence of *S. indicum* and *J. sambac* can be explained as a result of the proposed allopolyploidization at the base of Oleaceae, either by the loss of non-Oleacae Lamiales species or by recombination where the non-Oleaceae Lamiales copy was over-written (Figure S3.8). The other fraction of node D duplications with 4DTvs that map after the speciation of *J. sambac* are more difficult to explain, as in the trees they predate *J. sambac* divergence. This scenario is similar to the one we observe at the base of Oleaceae where topologically duplications are mapped at a different node than their age indicates. Therefore we propose that the Oleeae tribe was the result of a hybridization event with an ancestor in the lineage of *J. sambac* as one of the parents (Figure S3.8). In 1945 Taylor proposed that the Oleaceae group with 23 chromosomes (Oleoideae) had an allopolyploid origin whose ancestors were two probably extinct lineages from a group related to Jasminum with chromosome numbers of 11 and 12 (Taylor, 1945). This scenario is further supported by the use of a more stringent filtering of the trees (set2). When at least one sequence of *J. sambac* is in the clade, then the duplication density at node D increases from 0.37 to 0.63 (Figure S3.6). The use of a complete genome of *J. sambac* could confirm the allopolyploidization hypothesis at this point.

In order to confirm the two newly discovered allopolyploidization events with an alternative approach, we used GRAMPA (Gregg et al., 2017), which relies on gene-tree species-tree reconciliation to discern between allo- or auto-polyploidization. We performed two different analyses. In the first we compared the allopolyploidization model versus the autopolyploidization model at the base of Lamiales (node E) (see Figure S3.9a). We obtained lower parsimony scores for the allopolyploidization hypothesis (Table S3.2), indicating a better match with the gene trees as compared to an autopolyploidization scenario. We performed the same analysis comparing the proposed allopolyploidization at the base of the Oleeae lineage (node C) with two different hypotheses that place an autopolyplodization at the base of the family Oleaceae and at the base of the tribe Oleeae, respectively (see Figure S3.9b). The results once again supported allopolyploidization over each of the two autopolyplodization hypotheses. Finally, inspection of the phylome identified examples of gene trees that retained the duplications of the three polyploidization events, and whose topology is congruent with the proposed scenario (see Figure S3.10 as an example). Re-analysis of the syntenic depth results uncovered over 800 homologous syntenic regions with a depth of 8 between coffee and olive (see Figure 3.4).

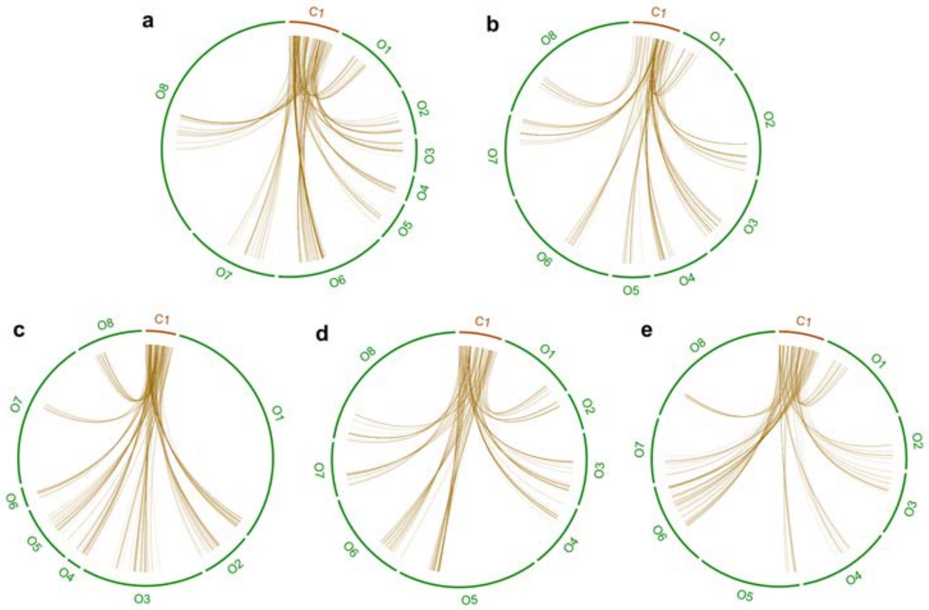

**Figure 3.4:** Example of five syntenic regions with a 1:8 relation between coffee and olive, as detected by GEvo.

## 3.4 Conclusions

Altogether our results underscore the power of phylogenomics to distinguish between allo- and auto-polyploidization. All our results indicate that the evolutionary history of olive comprises not only a species specific WGD, but also two older allopolyploidization events (Figure 3.5). The most ancestral event occurred at the base of the family Oleaceae, where a non-Oleaceae Lamiales species could be involved as one of the parental species. Also this event is independent of the one described before for the group of non-Oleaceae Lamiales species. The second one at the base of the Oleeae tribe that seems to involve a species related to Jasminum as one of the partners. The third event is specific to *O. europaea* and, with the current set of sequenced species, we do not find phylogenetic support for an allopolyploidization scenario. However, increased taxonomic sampling may change this.

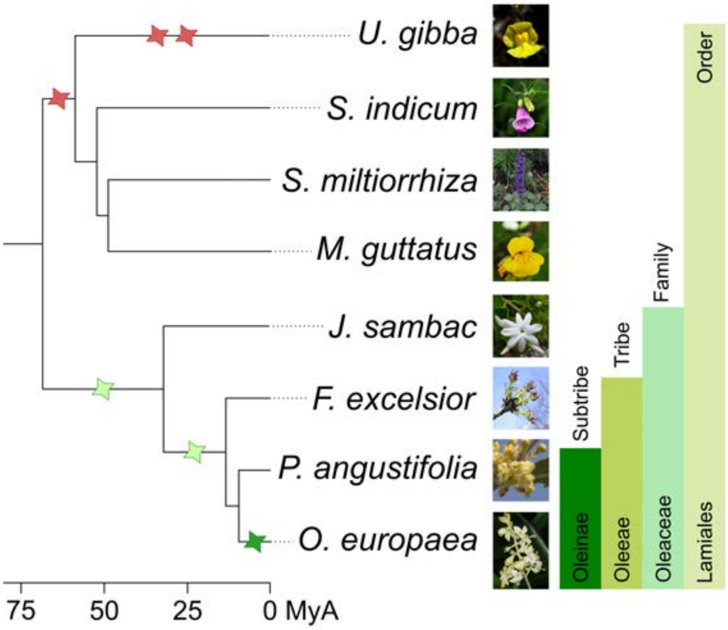

**Figure 3.5:** Species tree of the Lamiales clade showing the polyploidization events described in the literature (red stars) and in this analysis (green stars). The light green stars mark allopolyploidization events. Bars on the right show the taxonomic classification and the line in the bottom shows the divergence time in Mya.

## 3.5 Methods

### 3.5.1 Gene order analysis

The comparative genomic tools in the CoGe software package (Lyons et al., 2008) (https://genomevolution.org/coge/) were used to analyse gene order in the genomes of olive and its relatives. First, synmap was used to compare the olive genome against itself using the Syntenic Path Assembly option (Lyons et al., 2011) and removing scaffolds without conserved synteny (see Figure S3.1). Then, we used SynFind to obtain the syntenic depth, the number of conserved syntenic regions between one query genome and a reference. We obtained this value for comparisons of the olive, *Fraxinus excelsior* and *Sesamum indicum* using *Coffea canephora* as reference (see

Figure S3.1). SynFind was also used to find regions with a 1:8 relationship between coffee and olive (see Figure 3.4).

### 3.5.2 Phylome reconstruction

Six phylomes were reconstructed. In all cases an appropriate set of species was selected (see Table S3.1) and the PhylomeDB automated pipeline was used to reconstruct a tree starting from each gene encoded in each one of the seed genomes (Huerta-Cepas et al., 2011). This pipeline proceeds as follows: First a smith-waterman search is performed (Smith and Waterman, 1981) and the resulting hits are filtered based on the e-value and the overlap between query and hit sequences (e-value threshold <1e-05 and overlap >0.5). The filtered results are then aligned using three different methods (MUSCLE v3.8, MAFFT v6.814b and KALIGN 2.04) used in forward and reverse orientation (Edgar, 2004; Katoh et al., 2005; Lassmann and Sonnhammer, 2005; Landan and Graur, 2007). A consensus alignment is reconstructed from these alignments using M-coffee (Wallace et al., 2006). This consensus alignment is then trimmed twice, first using a consistency score (0.1667) and then using a gap threshold (0.1) as implemented in trimAl v1.4 (Capella-Gutiérrez et al., 2009). The resulting filtered alignment is subsequently used to reconstruct phylogenetic trees. In order to choose the best evolutionary model fitting each protein family, neighbor joining trees are reconstructed using BIONJ and their likelihoods are calculated using seven evolutionary models (JTT, WAG, MtREV, VT, LG, Blosum62, Dayhoff). The model best fitting the data according to the AIC criterion is then used to reconstruct a maximum likelihood tree with PhyML v3.1 (Guindon and Gascuel, 2003). All trees and alignments are stored and can be downloaded or browsed in phylomeDB (Huerta-Cepas et al., 2014) (http://phylomedb.org) with the Phylome IDs 215, 216, 217, 218, 219, and 220.

### 3.5.3 Incorporation of transcriptomic data in the olive phylome

Transcriptome data was downloaded from the ources indicated in their respective publications *Jasminum sambac* (Li et al., 2015b), and *Phillyrea angustifolia* (Sarah et al., 2017). In the case of *J. sambac*, where no protein prediction derived from the transcriptome was available, we obtained the longest ORF for each transcript. Only ORFs with a length of 100 aa or longer were kept, resulting in 20,952 ORFs in *J. sambac*. Transcriptomic data was introduced into each tree of the olive phylome using the following pipeline. First a similarity search using blastP was performed from the seed protein against a database that contained the two transcriptomes. Results were then filtered based on three thresholds: e-value <1e-05, overlap between query and hit had to be at least of 0.3, and a sequence identity threshold >40.0%. Hits that passed these filters were incorporated into the raw alignment of the phylome using MAFFT (v 7.222) ( - -add and - -reorder options) (Katoh and Frith, 2012). Then trees were reconstructed using the resulting alignment and following the same procedure as described above. Once all trees were reconstructed, they were filtered to remove unreliably placed transcriptome sequences. Phylomes tend to be highly redundant, specially when the seed genome contains many duplications, as is the case for the olive genome. Therefore, the same transcriptomic sequence is likely inserted in many trees. For each inserted transcript, we checked whether the sister sequences of each inserted transcript overlapped. If such overlap did not exist the transcript was deemed unreliable and removed from the tree. This filtered set was then filtered once more to select trees that contained at least one transcript for each of the two new species (set1). Finally set1 was filtered again to keep only trees that contained a monophyletic clade including the four Oleaceae species (set2).

### 3.5.4 Species tree reconstruction

A species tree was reconstructed using data from the olive phylome. Each tree reconstructed for this phylome was first pruned so that species specific duplications were deleted from the tree, keeping only one sequence as representative of the duplicated group. Once trees were pruned, only those trees that contained one sequence for each of the 19 species included in the phylome were selected. 215 such trees were found. The clean alignments used to reconstruct these trees were concatenated and a species tree was reconstructed using the model of amino acids substitution LG implemented in PhyML v3.1 (Guindon and Gascuel, 2003) with 100 bootstrap replicates. In addition, a second species tree was reconstructed using a super-tree approach with the tool duptree (Wehe et al., 2008). In this case all trees in the olive phylome were used for the tree reconstruction. A third species tree was reconstructed after the inclusion of the transcriptomic data into the olive phylome. From the initial set of genes chosen to reconstruct the first species tree, a subset was chosen to reconstruct the extended species tree. This subset included only genes that incorporated at least one of the three species with a transcriptome. This final tree was reconstructed using 112 gene alignments using the same methodology as described above.

### 3.5.5 Detection and mapping of orthologs and paralogs

Orthologs and paralogs were detected using the species overlap method (Huerta-Cepas et al., 2007) as implemented in ETE v3.0 (Huerta-Cepas et al., 2016). Species specific duplications (expansions) are computed as duplications that map only to one species, in our case always the species from which the phylome was started. In order to reduce the redundancy in the prediction of species specific expansions a clustering is performed in which expansions that overlap in more than 50% of their sequences are fused together. Predicted duplication nodes are then mapped to the species tree under the assumption that the duplication happened at the common ancestor of all the species included in the node, as described by Huerta-Cepas and Gabaldón (Huerta-Cepas et al., 2011). Duplication frequencies at each node in the species tree are calculated by dividing the number of duplications mapped to a given node in the species tree by all the trees that contain that node. In all cases duplication frequencies are calculated excluding trees that contained large species specific expansions (expansions that contained more than 5 members).

### 3.5.6 GO term enrichment

GO terms were assigned to the olive proteome using interproscan (Jones et al., 2014) and the annotation of orthologs from the phylomeDB database (Huerta-Cepas et al., 2014). Phylome annotations were transferred to the olive proteome using one-to-one and one-to-many orthologs. GO term enrichment of proteins duplicated at the different species-specific expansions and duplication peaks was calculated using FatiGO (Al-Shahrour et al., 2004).

### 3.5.7 Topological analysis

A topological analysis was performed using ETE v3.0 (Huerta-Cepas et al., 2016) to test whether a duplication event happened at the base of Lamiales and determine which species were involved. We searched how many trees supported each of the following topologies: the complete topology where at least one Oleaceae and at least one other non-Oleaceae Lamiales are found at both sides of the duplication (topology TA), a partial topology where all non-Oleaceae Lamiales species have been lost in one side of the duplication (topology TB), or another partial topology where the Oleaceae sequences have been lost at one side of the duplication (topology TC) (see Figure 3.2a). The analysis was then repeated in different previously reconstructed phylomes that contained ancient whole genome duplications where there was an imbalance of species at either side of the duplication. The phylomes selected were those of the plants *Phaseolus vulgaris* (Vlasova et al., 2016) (Phylome ID 8) and *Solanum commersonii* (Aversano et al., 2015) (Phylome ID 147), the fish *Scophthalmus maximus* (Figueras et al., 2016) (Phylome ID 18), and the fungi *Rhizopus delemar* (Corrochano et al., 2016) (Phylome ID 252). Each of those phylomes contains an old WGD where at one side of the duplication there are less species than at the other one. We checked the proportion of trees that supported each topology. Like with the Oleaceae example, topology TA′ conserves at least one member of each group, topology TB′ has lost all the species of the large group at one side of the duplication while TC′ has lost all the species of the small group at one side of the duplication (see Figure 3.2d). We used GRAMPA (Gregg et al., 2017) (Spring 2016 version) to assess five different hypothesis (see Figure S3.10) using the two sets of trees that contained transcriptomic data. This tool uses reconciliation in order to compute the support between a set of trees and a proposed allopolyploidization or autopolyploidization event. Though it is limited to detecting one single event at a time. During its calculation, GRAMPA discards single gene trees that have too many possibilities when reconciling them to the species tree. The trees discarded can vary depending on the species tree hypothesis. Therefore, in order to fairly compare the parsimony scores obtained, we recalculated them based on the trees used in all the hypotheses. We performed two different analyses. In the first we compared the allopolyploidization model versus the autopolyploidization at the base of Lamiales (see Figure S3.10a). In the second we compared the allopolyploidization that led to the Oleeae lineage with two different hypotheses that place an autopolyploidization at the base of Oleaceae family and at the base of Oleeae tribe respectively (see Figure S3.10b). Results can be found in Table S3.2.

### 3.5.8 Transversion rate at fourfold degenerate sites (4DTv)

The 4DTv distribution was used to estimate speciation and polyploidization events. In order to obtain the gene pairs we used the species tree that included the transcriptomic data. We calculated the 4DTv values for the orthologous gene pairs between *O. europaea* with *J. sambac*, *F. excelsior*, *P. angustifolia*, and *S. indicum*. We also calculated the 4DTv for each paralogous gene pair of olive that maps at each evolutionary age.

### 3.5.9 Divergence times

Divergence times were calculated using r8s-PL 1.81 (Sanderson, 2003). Four nodes were taken as calibration points. The divergence time of these nodes were obtained from the TimeTree database (Hedges et al., 2015): *Mimulus guttatus - Arabidopsis thaliana* (117 Mya), *Sesamum indicum - Solanum lycopersicum* (84 Mya), *Glycine max - Arabidopsis thaliana* (106 Mya), *Zea mays - Solanum lycopersicum* (160 Mya). Cross-validation was performed to choose the smoothing parameter.

### Acknowledgements

TG group acknowledges support from the Spanish Ministry of Economy and Competitiveness grants, ‘Centro de Excelencia Severo Ochoa 2013-2017’ SEV-2012-0208, and BFU2015-67107 cofounded by European Regional Development Fund (ERDF); from the European Union and ERC Seventh Framework Programme (FP7/2007-2013) under grant agreements FP7-PEOPLE-2013-ITN-606786 “ImResFun” and ERC-2012-StG-310325; from the Catalan Research Agency (AGAUR) SGR857, and grant from the European Union’s Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No H2020-MSCA-ITN-2014-642095; TG and PV acknowledge support from Banco Santander for the olive genome sequencing project; IJ was supported in part by a grant from the Peruvian Ministry of Education: 'Beca Presidente de la República' (2013-III).

## 3.A Supplementary material

**Table S3.1:** List of species included in the reconstruction of the six phylomes used in this study. Columns indicate, in this order, the species code for each species, the species name, the source for the protein and the coding DNA sequences, and the phylome in which the species was used (*O. europaea*-215, *F. excelsior*-216, *M. guttatus*-217, *S. indicum*-218, *U. gibba*-219, *S. miltiorrhiza*-220).

| Species code | Species name | Source of protein coding sequences | PhylomeId |
|---|---|---|---|
| OLEEU | *Olea europaea* | *Olea europaea* genome project | 215, 216, 217, 218, 219, 220 |
| FRAEX | *Fraxinus excelsior* | http://www.ashgenome.org | 215, 216, 217, 218, 219, 220 |
| UTRGI | *Utricularia gibba* | *Utricularia gibba* Genome Sequencing Project | 215, 216, 217, 218, 219, 220 |
| SESIN | *Sesamum indicum* | ocri-genomics.org | 215, 216, 217, 218, 219, 220 |
| MIMGU | *Mimulus guttatus* | JGI | 215, 216, 217, 218, 219, 220 |
| SALMI | *Salvia miltiorrhiza* | Chinese Herbal Plant Genome Database | 215, 216, 217, 218, 219, 220 |
| COFCA | *Coffea canephora* | coffee-genome.org | 215, 216, 217, 218, 219, 220 |
| HELAN | *Helianthus annuus* | biodiversity.ubc.ca | 215, 216, 217, 218, 219, 220 |
| SOLLC | *Solanum lycopersicum* | ENSEMBL | 215, 216, 217, 218, 219, 220 |
| ARATH | *Arabidopsis thaliana* | Ensembl Plants - Release 17 | 215, 216, 217, 218, 219, 220 |
| AMBTC | *Amborella trichopoda* | Uniprot | 215, 216, 217, 218, 219, 220 |
| BETVU | *Beta vulgaris* | CRG Ultrasequencing Unit | 215 |
| VITVI | *Vitis vinifera* | Ensembl Plants release 25 | 215 |
| POPTR | *Populus trichocarpa* | EnsemblPlants - Release 15 | 215 |
| SOYBN | *Glycine max* | Ensembl Plants - Release 17 | 215 |
| ORYSJ | *Oryza sativa* subsp. *japonica* | Ensembl Plants - Release 22 | 215 |
| TOBAC | *Nicotiana tabacum* | Solgenomics | 215 |
| SOLTU | *Solanum tuberosum* | Ensembl Plants - Release 22 | 215 |
| MAIZE | *Zea mays* | Ensembl Plants release 25 | 215 |

**Table S3.2:** List of parsimony scores for each of the different hypothesis and considering the two sets of trees with EST data. Nodes are named as shown in Figure 3.3.

| Event | Hypothesis | Parsimony score for set1 | Parsimony score for set2 |
|---|---|---|---|
| Event Oleaceae | Hybridization at node 5 | 663887 | 440499 |
| | WGD at node 5 | 669725 | 443697 |
| Event Oleeae | Hybridization at node 3 | 663379 | 428200 |
| | WGD at node 3 | 664593 | 440452 |
| | WGD at node 4 | 666795 | 437880 |

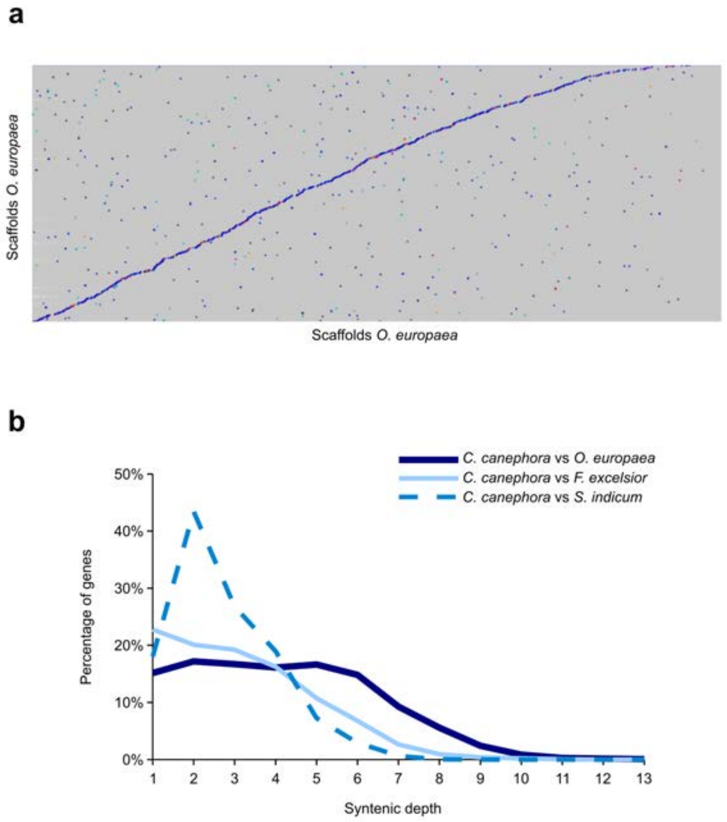

**Figure S3.1:** Results obtained with the GOGE package. a) Image of a mapping of *O. europaea* against itself as shown by Synmap. b) Syntenic depth as calculated by SynFind.

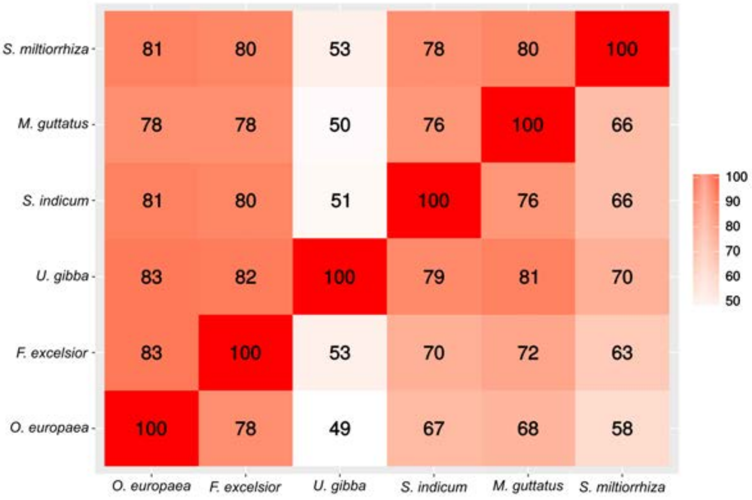

**Figure S3.2:** Heatmap showing the percentage of orthologous proteins in comparison to each Lamiales species included in this analysis.

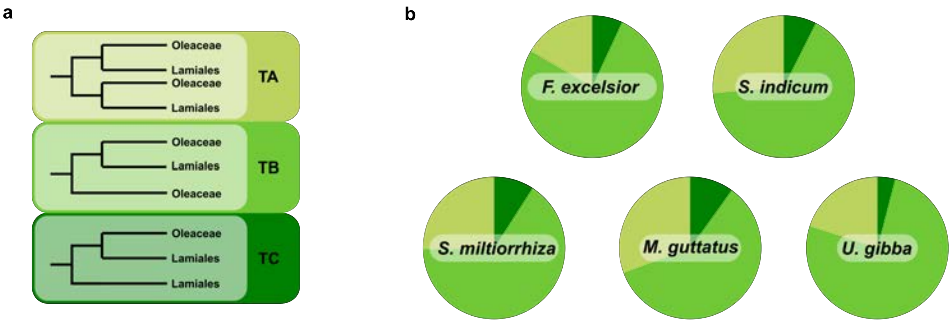

**Figure S3.3:** Pie-charts representing the distribution of trees supporting each of the topologies as shown in Figure 3.2

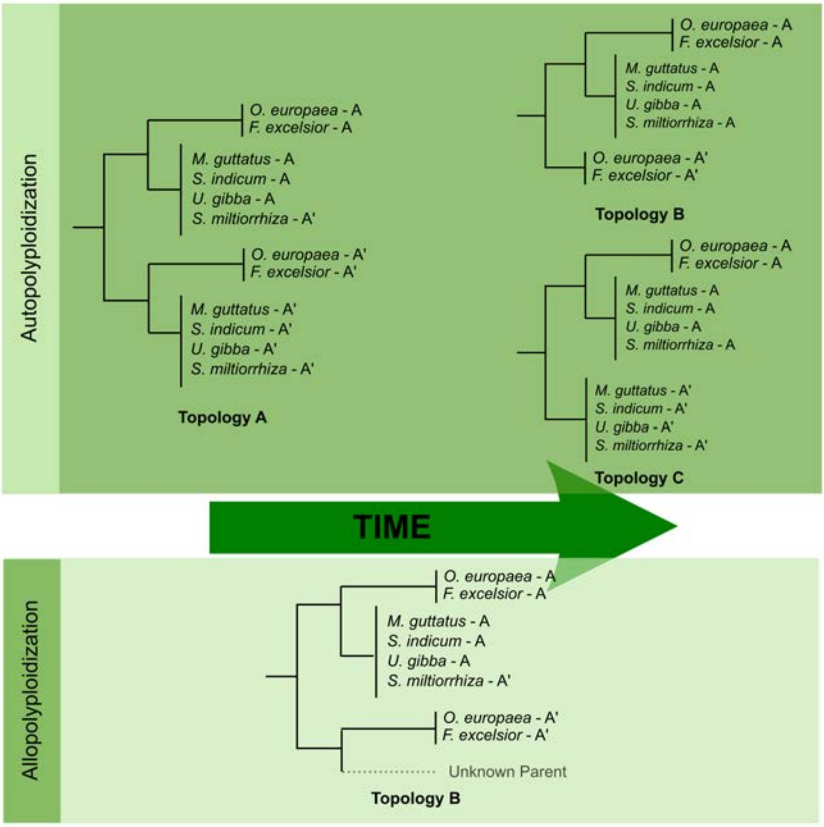

**Figure S3.4:** Exact topologies expected to find in a scenario of autopolyploidization and one of allopolyploidization.

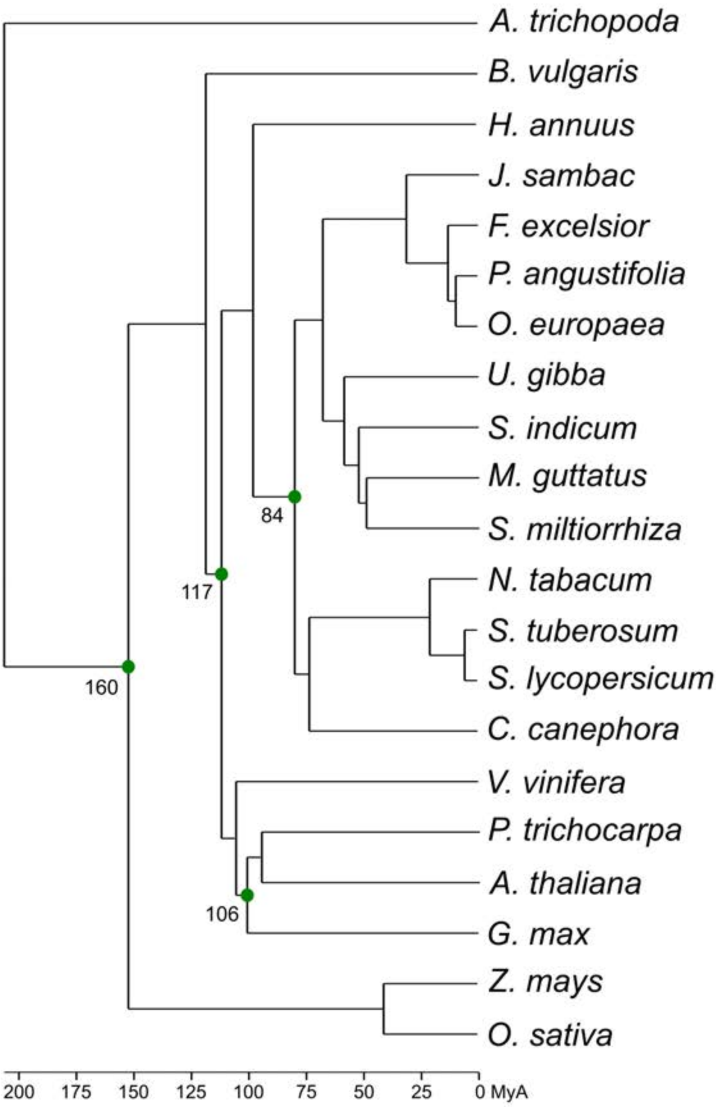

**Figure S3.5:** Chronogram depicting the evolution of the plants included in the phylome. Green dots represent selected calibration points in Mya.

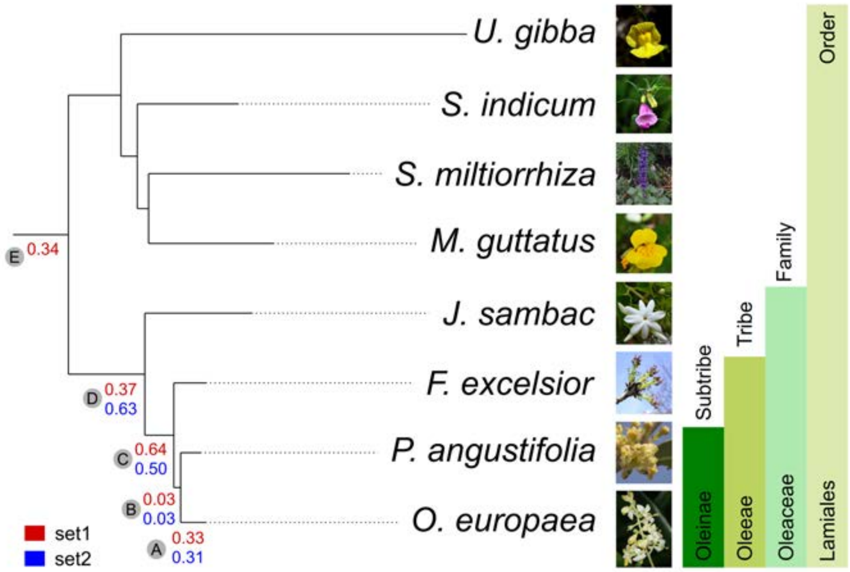

**Figure S3.6:** Species tree of the Lamiales order, including *P. angustifolia*, *F. excelsior* and *J. sambac*. The duplication rates are shown in red for set1 and in blue for set2. The grey circles show the node name and the bars on the right, the taxonomic classification.

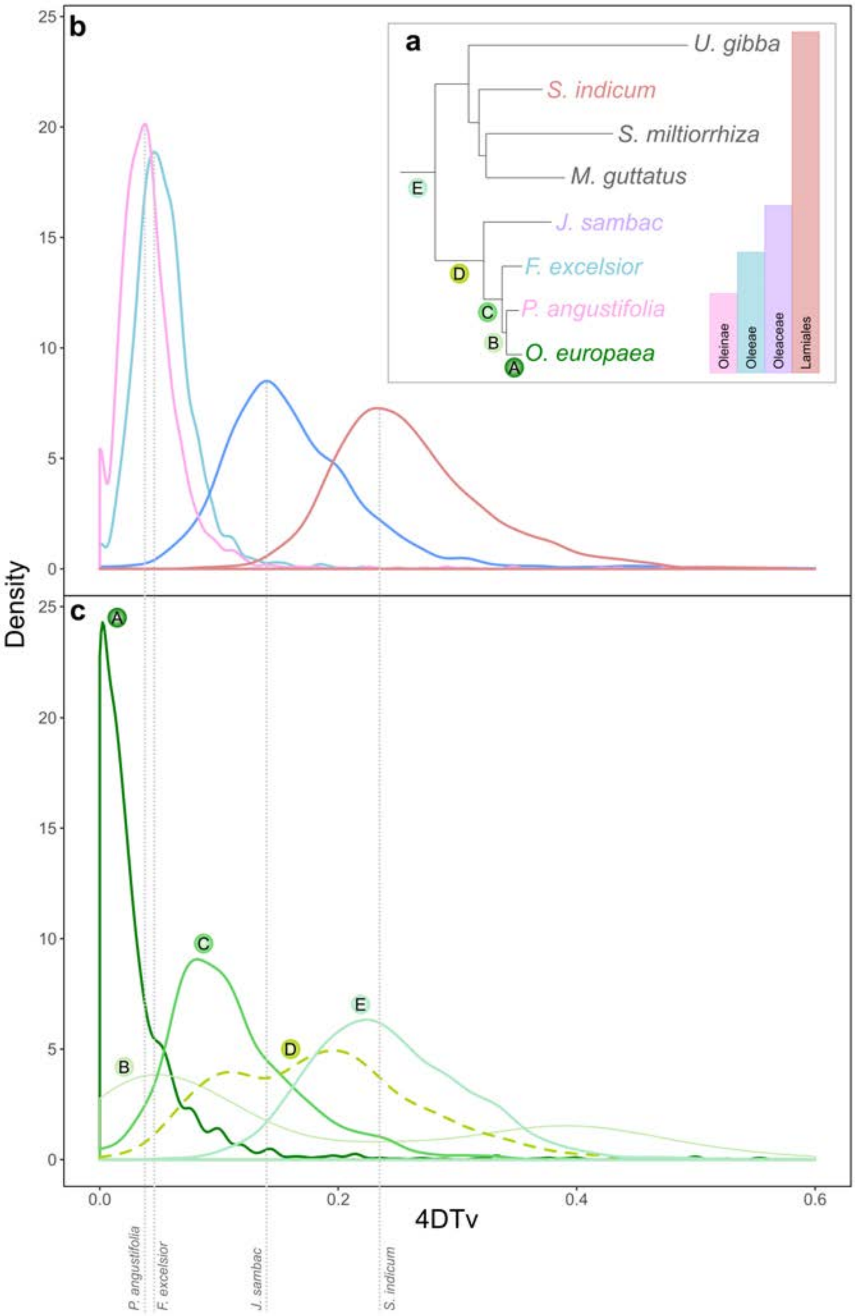

**Figure S3.7:** Species tree and 4DTv of the set2. a) species tree of the group of Lamiales including the three Oleaceae species. Nodes where the 4DTv of the paralogous pairs were calculated are marked with letters (A to E) as referred to in the text and coloured according each evolutionary age. The species used to calculate the 4DTv of orthologous pairs are shown in different colours. The bars on the right show the taxonomic classification. b) 4DTv of the orthologous pairs between *O. europaea* with *P. angustifolia*, *F. excelsior*, *J. sambac* and *S. indicum*. c) 4DTv of the paralogous pairs of *O. europaea* at the marked nodes in the tree.

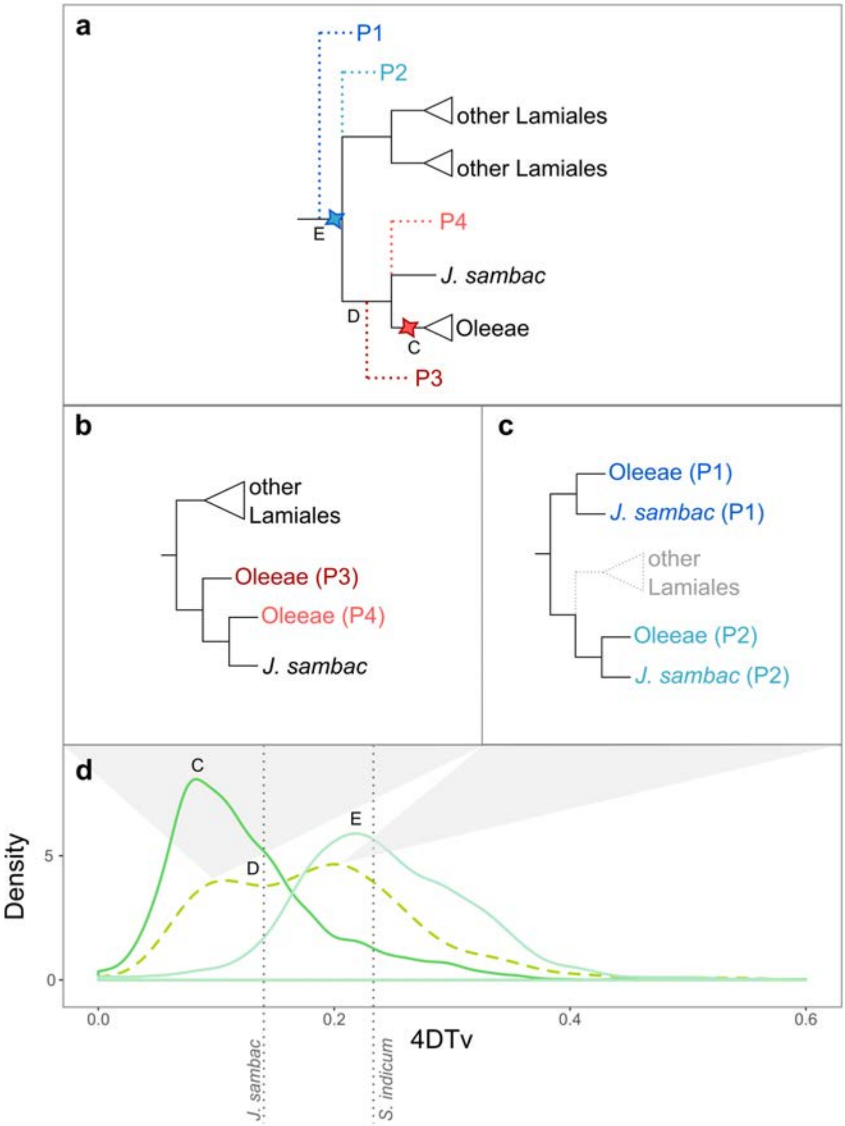

**Figure S3.8:** Schematic explanation of the 4DTv density at node D in Figure 3.3c. a) representation of the two allopolyploidization events and the potential parentals.b) scheme of a gene tree where the protein of *J. sambac* map after the divergence of this species. c) scheme of a gene tree where the non-Oleaceae Lamiales proteins are lost. d) 4DTv of the paralogs at nodes C, D, and E. The dotted lines mark the divergence time between olive - *J. Sambac* and olive - *S. indicum*.

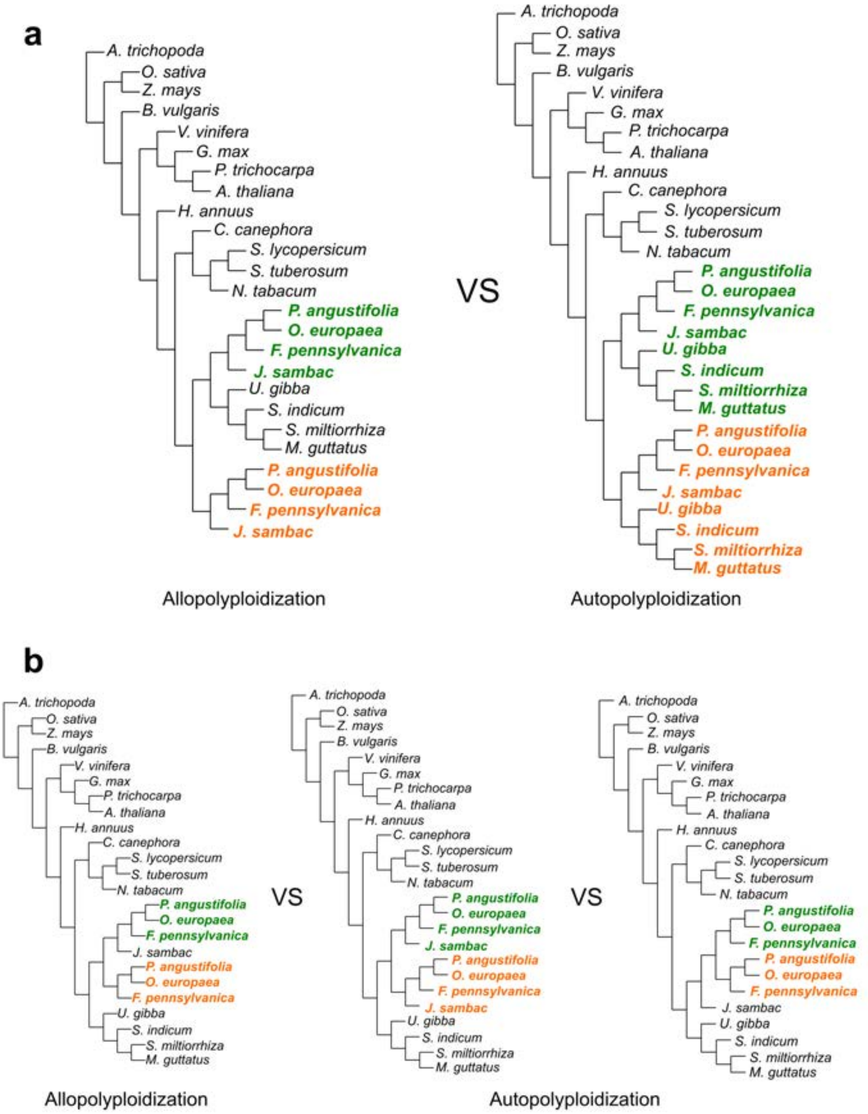

**Figure S3.9:** Phylogenetic trees representing the comparisons done for GRAMPA. In all cases branches painted in green and orange represent the species that the polyploidy has affected. a) The trees represent the hypothesis of an allopolyploidization versus an autopolyploidization at the base of Lamiales. b) These trees represent the hypothesis of an allopolyploidization versus a two models of autopolyploidization.

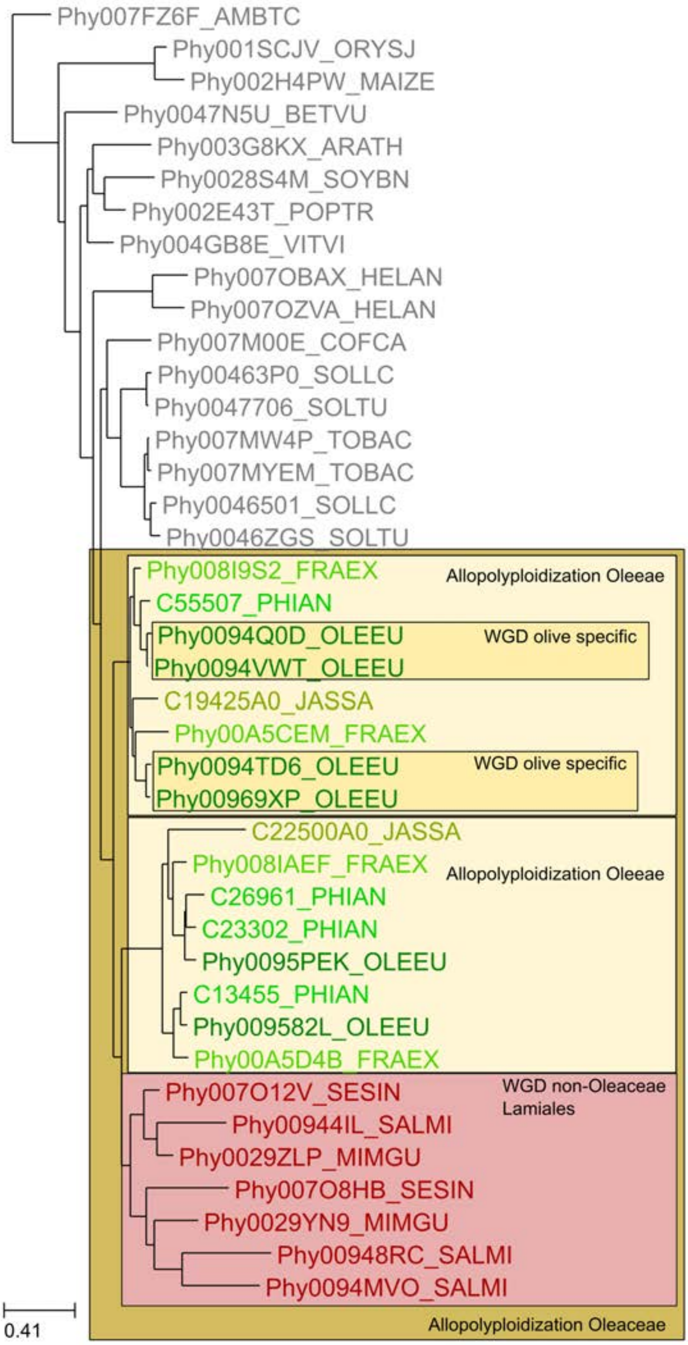

**Figure S3.10:** Example gene tree that shows the three events we have described in olive: the species specific duplication and the two allopolyploidizations. The whole genome duplication previously described in non-Oleaceae Lamiales and the species specific duplications in *U. gibba* can also be seen.

# 4 Genome analysis of the *Olea europaea* complex

## Genome sequencing of wild and cultivated olive trees reveals rampant hybridization in the *Olea europaea* complex

This chapter provides the first phylogenomic analysis of the *O. europaea* complex.

The *Olea europaea* complex is composed by six subspecies: *europaea*, *cerasiformis*, *maroccana*, *guanchica*, *laperrinei*, and *cuspidata*. The Mediterranean olives belong to the subsp. *europaea*, which is divided in two varieties: the oleasters (var. *sylvestris*) and the cultivars (var. *europaea*). In this study, we sequenced and annotated the plastid and the mitochondrial genomes of the same individual (*O. europaea* cv. 'Farga') used for the sequencing of the nuclear genome (chapter 2). In addition, we sequenced the genomes of twelve additional individuals covering morphological diversity, taxonomy and geographical distribution: five of subsp. *europaea* (four cultivars, one oleaster), three of subsp. *cuspidata*, one of *cerasiformis*, one of *maroccana*, one of *guanchica* and one of *laperrinei*. Despite the consideration of a diploid status for the cultivated olive (subsp. *europaea*), the analysis of allele depth frequency showed that all the individuals of the *O. europaea* complex underwent a more recent polyploidization, ancestral to the divergence of all the subspecies. This analysis complement the results of the chapter 3.

Comparative analysis of all the individuals of the *O. europaea* complex showed a high nucleotide diversity (nuclear, plastid, and mitochondrial) of the wild individuals (oleaster plus the other five subspecies) in comparison with cultivars. A selection test showed that some proteins associated with stress response and developmental processes are positively selected in cultivars.

In addition, many studies have tried to understand the phylogenetic relationships between these individuals using genetic markers. In this project we decided to perform a comprehensive analysis using whole nuclear and organellar genome data. The prediction of SNPs for all available genomes was used in order to reconstruct phylogenetic relationships among the *O. europaea* complex. Our results showed that the history of the *O. europaea* complex has experienced rampant admixture of genetic material between all the subspecies thru gene flow. Moreover, our results support a former hypothesis point that var. *sylvestris* is not monophyletic. On the contrary, it shows evidence of a continuous hybridization process between wild and cultivated olive trees. Finally, the evolutionary history of the cv. 'Farga' reveals that it may be different from the other studied cultivars. An ancestral hybridization event is suggested in the formation of this cultivar, involving an individual already domesticated in the east of the Mediterranean basin and an individual of var. *sylvestris* from the west. All these results are further supported by the inconsistencies observed between the nuclear and the plastid phylogenetic trees.

Irene Julca, Marina Marcet-Houben, Fernando Cruz, Ivo G. Gut, Tyler S. Alioto, Pablo Vargas, and Toni Gabaldón. Genome sequencing of wild and cultivated olive trees reveals rampant hybridization in the *Olea europaea* complex. *In preparation*.

# Genome sequencing of wild and cultivated olive trees reveals rampant hybridization in the *Olea europaea* complex

## 4.1 Abstract

The olive tree (*Olea europaea* L.) is an iconic species of the Mediterranean basin. It is subdivided into six subspecies (*europaea*, *laperrinei*, *guanchica*, *maroccana*, *cerasiformis*, and *cuspidata*), and two botanical varieties are considered within the subsp. *europaea*: the cultivated forms (var. *europaea*) and the wild types (var. *sylvestris*). Tracing the recent evolution and genetic diversification of the species is paramount for the management and preservation of genetic resources, and for understanding the key process of olive tree domestication. In order to study the recent history of the *O. europaea* complex, we sequenced the genomes of four cultivars ('Arbequina', 'Beladi', 'Picual' and 'Sorani'), one wild type, and at least one individual of each of the other subspecies. Altogether, twelve whole genomes, including the recently sequenced genome of the cv. 'Farga', were analysed. Our results reveal high, but varying levels of heterozygosity in both cultivated and wild relatives. Genes specifically under selection in cultivars include many genes associated with the biotic and abiotic stress response, among other physiological and developmental processes. Notably, the patterns of relative coverage of alternative alleles suggest the existence of a relatively recent polyploidization that is shared by all subspecies. Phylogenomic reconstruction based on whole nuclear and organellar genomic data reveals a network-like diversification, with evidence of large levels of genetic admixture and phylogenetic incongruence. Indeed, we detect extensive gene flow between wild and cultivated trees, including recent admixture of the eastern and western Mediterranean populations in the case of cv. 'Farga'. In a nutshell, our results shed light on the recent evolution of the *O. europaea* complex, highlighting the plasticity of its genome and the impact of hybridization and genome duplication in the evolutionary history of the olive tree.

## 4.2 Introduction

The Mediterranean olive tree (*Olea europaea* L. subsp. *europaea*) is one of the earliest cultivated fruit trees of the Mediterranean basin. Archaeological, palaeobotanical, and genetic studies situate the first evidence for olive cultivation around 6,000 years ago in the eastern Mediterranean basin (Zohary and Spiegel-Roy, 1975; Kaniewski et al., 2012; Terral et al., 2004; Besnard et al., 2013b). However, it is still unclear whether cultivated trees derived from a single initial domestication event in the Levant, followed by secondary diversification (Besnard et al., 2013b; Besnard and Rubio de Casas, 2016), or whether cultivated lineages are the result of more than one independent domestication event (Diez et al., 2015; Breton et al., 2009; Díez and Gaut, 2016; Terral et al., 2004; Yoruk and Taskin, 2014). The olive tree (*O. europaea*, Oleaceae) is divided into six subspecies, which collectively are referred to as the *O. europaea* complex and include: *europaea*, *laperrinei*, *guanchica*, *maroccana*, *cerasiformis*, and *cuspidata* (Green, 2002; Vargas et al., 2000). The subsp. *europaea* is further subdivided into two taxonomic varieties: var. *sylvestris*, also named oleaster, which encompasses the wild forms of the olive tree, and var. *europaea*, which comprises cultivated forms (Green, 2002). Recent analyses enabled by the sequencing of the first complete genome of *O. europaea* have uncovered several ancient polyploidization events in the lineage leading to this species, of which two were described as allopolyploidization events (Julca et al., 2017; Cruz et al., 2016a). In addition to these ancient events, earlier work had described hybridization processes in the *O. europaea* complex (Besnard et al., 2007b, 2009; Rubio de Casas et al., 2006) and had identified the presence of a polyploid series, including the so-called diploid subspecies (*europaea*, *laperrinei*, *cuspidata*, *guanchica*) with 2n = 46, the tetraploid subsp. *cerasiformis* and the hexaploid subsp. *maroccana* (Green and Wickens, 1989; Besnard et al., 2008). However, these studies are limited to the use of a reduced number of genetic markers and chromosome counting techniques. Access to whole genome sequences of additional cultivars and subspecies of the *O. europaea* complex is necessary to understand the recent evolution of this plant, and to assess the genomic aftermath of past processes of hybridization and whole genome duplication.

To gain insight into the recent evolution of the *O. europaea* complex we sequenced twelve trees, including four cultivated individuals of the var. *europaea* (cultivars: 'Arbequina', 'Beladi', 'Picual' and 'Sorani'), one wild individual of var. *sylvestris*, and at least one individual from each of the other five subspecies of the *O. europaea* complex (*laperrinei*, *cuspidata*, *guanchica*, *maroccana* and *cerasiformis*). The analysis of genome sequences from these twelve individuals and its comparison with the available reference genome (cultivar 'Farga') (Cruz et al., 2016a) may shed light on the recent evolution and domestication of this species. Our results revealed a higher nucleotide diversity in wild individuals, as compared to cultivated ones. We found that, in cultivated trees, genes associated with stress response and development processes were common among those predicted to be under positive selection. Furthermore, patterns of allelic coverage at heterozygous sites showed that a significant part of the genomes of all supposedly diploid individuals presented signatures of tetraploidy, suggesting a relatively recent polyploidization prior to the divergence of the different subspecies. Finally, phylogenomic analyses using genetic polymorphisms over the entire organellar and nuclear genomes revealed that hybridization has been an important process in the evolution of the *O. europaea* complex. Moreover, hybridization between the different varieties (*europaea* and *sylvestris*) of subsp. *europaea* has been common. Particularly, the cultivar 'Farga' has a different maternal origin as compared to the other studied cultivars. In sum, our results highlight the high genomic plasticity of the olive tree genome and underscore the impact of hybridization within the *O. europaea* complex.

## 4.3 Results and discussion

### 4.3.1 Patterns of genetic polymorphism in *O. europaea*

In order to analyze the genetic diversity and recent evolution of the olive tree we sequenced twelve individuals, including at least one sample per each of the defined subspecies of the *O. europaea* complex, collected in different geographical regions that represent the current distribution of the species (see Table S4.1). SNPs were called at the nuclear, plastid and mitochondrial genome, using the respective sequence of the cultivar 'Farga' as a reference (Cruz et al., 2016a) (see material and methods). Altogether we obtained a total of 18,399,785 polymorphic positions uniformly distributed along the nuclear genome (see Figure S4.1), 214 in the plastid genome, and 2,561 in the mitochondrial genome (see Figure S4.2). In the plastid, a large region ($\sim$25 Kb) is fully conserved and devoid of SNPs in all analyzed individuals (Figure S4.3, Figure S4.2). Interestingly, this region contains the largest plastid gene, *ycf2*, which also has a low rate of nucleotide substitution in other plants (Huang et al., 2010). This gene is essential for plant survival, however the exact function is unknown (Drescher et al., 2000; Wicke et al., 2011). The conserved region also comprises other genes such as *ycf15*, *rps7*, *rps12*, *ndhB*, rRNA and tRNA genes.

The general nucleotide diversity (Nei's index, Hs) in the *O. europaea* complex was 3.3x10$^{-3}$ for the nucleus, 0.4x10$^{-3}$ for the plastid, and 1.1x10$^{-3}$ for the mitochondria. If we compare the nucleotide diversity of the cultivars with that of the wild forms (oleaster plus other subspecies), we can see a reduction of the Hs in the cultivars (see Table 4.1). This pattern was previously observed in other studies based on ISSRs (Vargas and Kadereit, 2001), allozyme polymorphisms (Lumaret et al., 2004), SSRs (Belaj et al., 2010), and plastid DNA variations (Besnard et al., 2011). In general, lower genetic diversity in cultivars is commonly associated with genetic bottlenecks during domestication (Doebley et al., 2006). Such reduction in genetic diversity has also been observed in other cultivated plants such as *Prunus persica* (International Peach Genome Initiative et al., 2013), *Vitis vinifera* (Zhou et al., 2017), *Pyrus ussuriensis* (Cao et al., 2012), *Malus domestica* (Zhang et al., 2012b; Velasco et al., 2010), citrus species (Wang et al., 2017b), cucumber (Qi et al., 2013), tomato (Sauvage et al., 2017), among others.

**Table 4.1:** Nei's index of the cultivated olives and the wild individuals (oleaster plus other subspecies) for the nuclear, plastid and mitochondrial sequences.

| | Nucleous | Plastid | Mitochondria |
|---|---|---|---|
| cultivars | 1.86 x $10^{-3}$ | 0.15 x $10^{-3}$ | 0.82 x $10^{-3}$ |
| wild individuals | 3.44 x $10^{-3}$ | 0.45 x $10^{-3}$ | 0.97 x $10^{-3}$ |

Expectedly, the amount of nuclear polymorphisms of the different individuals with respect to the reference, correlated with their known evolutionary distances so that within the subsp. europaea, var. *sylvestris* has slightly more SNPs (4.18 SNPs/Kb) than the cultivars (~3.32 SNPs/Kb) (Figure 4.1a), whereas individuals of the other subspecies had a larger number of SNPs, with the subsp. *cuspidata* from Iran showing the highest number (9.11 SNPs/Kb). In addition, the number of heterozygous SNPs was high in the two subspecies that entail polyploidy: *maroccana* is described as hexaploid and *cerasiformis* as tetraploid (Besnard et al., 2008).

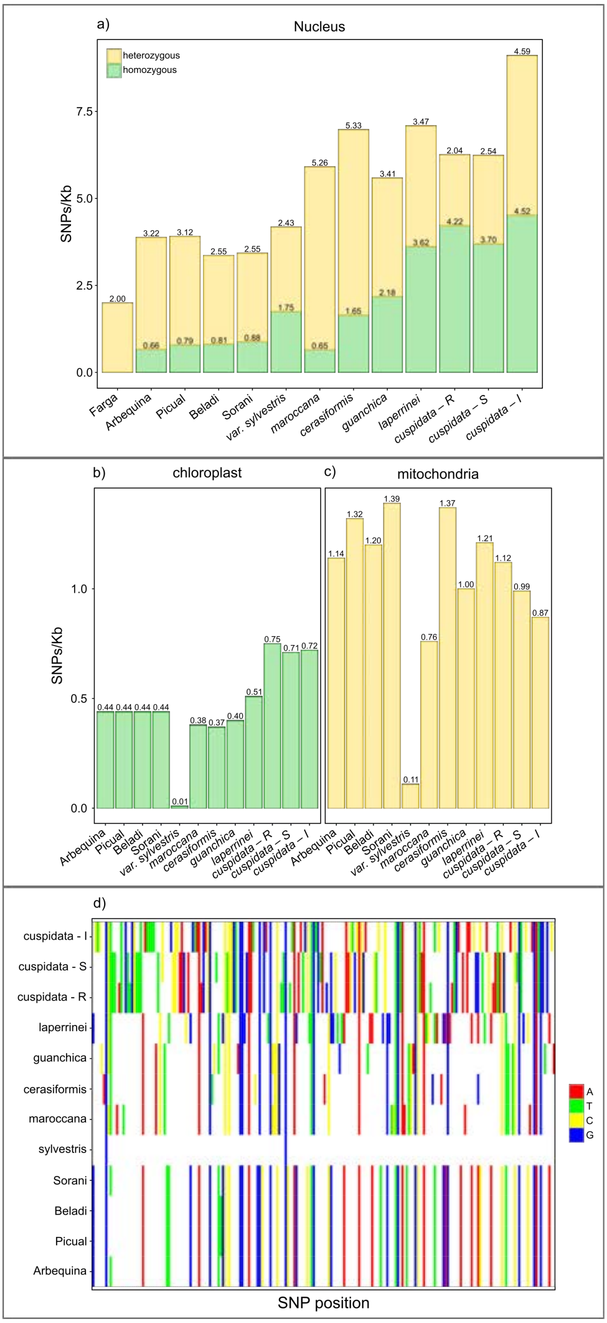

**Figure 4.1:** SNP densities (SNPs/Kb) in sequenced individuals, densities for homozygous and heterozygous SNPs are indicated separately. Densities are indicated for the nuclear (a), plastid (b), and mitochondrial (c) genomes. d) Plot showing the relative position and identity of plastid SNPs as compared to the cv. ‘Farga’ reference.

Strikingly, the patterns of polymorphisms in the organellar genomes did not follow the gradient described above for the subsp. *europaea* (Figure 4.1b). In contrast to the nuclear genome, the plastid and mitochondrial genomes of the wild individual var. *sylvestris* show a significant lower number of SNPs (see Figure 4.1b,c). Specifically for the plastid we can observe that cultivars of var. *europaea* share 61 polymorphic positions when compared to 'Farga', while the var. *sylvestris* only shows two SNPs (Figure 4.1d). This clearly indicates that nuclear and organellar genomes tell different evolutionary stories for our reference genome. In addition, from our results we can conclude that the organellar genomes of the cv. 'Farga', used as reference, and the wild individual var. *sylvestris*, sequenced in this study, show a very close genetic relationship. As organelles are maternally inherited in olive (Besnard et al., 2000), our results suggest that the maternal lineage of the cv. 'Farga' derives from wild individuals from oleaster populations of the western Mediterranean basin (represented by the individual sequenced here). In contrast, all the other cultivars share a very distinct organellar haplotype, likely derived from oleaster populations from the eastern Mediterranean, not represented here (but see phylogenetic analyses of plastid genomes below). Evolutionary relationships among sequenced individuals are further investigated below.

### 4.3.2 Identification of genes selected in the domestication of olives

Patterns of polymorphism can serve to detect genes under selection. In order to search for genes putatively under positive selection in the cultivars, we classified the SNPs into intergenic, intronic, and coding. We further classified coding SNPs according to whether they imply synonymous or nonsynonymous changes (see material and methods). As we can see in Table 4.2 and Figure 4.2, in all the cases a higher percentage of SNPs are present in intergenic regions (4.1 SNPs/Kb), followed by the intronic region (0.8 SNPs/Kb). In coding regions only 0.3 SNPs/kb are present, and the number of synonymous and nonsynonymous changes is similar across accessions. In order to assess selection in the *O. europaea* complex we first measured the ratio of nonsynonymous and synonymous nucleotide diversity ($\pi N/\pi S$) in all the sequenced genomes included in this study. For all the cases the $\pi N/\pi S$ was similar, with an average of 0.38 (Table 4.2), suggesting similar strengths of selective pressure across all the genomes. This ratio is similar to that found for other trees such as *Populus nigra* (0.48) (Chu et al., 2009) and *Populus trichocarpa* (0.40) (Tuskan et al., 2006).

**Table 4.2:** Number of synonymous and nonsynonymous homozygous SNPs, and synonymous and nonsynonymous heterozygous SNPs per individual. The columns number four, seven and eight show the $\pi N/\pi S$ ratio of the homozygous, heterozygous, and total number of SNPs, respectively.

| *O. europaea* subsp. | Homo<br>Syn | Homo<br>Non-syn | Homo<br>$\pi N/\pi S$ | Hetero<br>Syn | Hetero<br>Non-syn | Hetero<br>$\pi N/\pi S$ | Total<br>$\pi N/\pi S$ |
|---|---|---|---|---|---|---|---|
| var. *europaea* 'Farga' | 0.00 | 0.00 | 0.00 | 0.05 | 0.06 | 0.37 | 0.37 |
| var. *europaea* 'Arbequina' | 0.02 | 0.02 | 0.37 | 0.08 | 0.09 | 0.38 | 0.38 |
| var. *europaea* 'Picual' | 0.02 | 0.02 | 0.37 | 0.08 | 0.09 | 0.38 | 0.38 |
| var. *europaea* 'Beladi' | 0.02 | 0.03 | 0.37 | 0.06 | 0.07 | 0.39 | 0.38 |
| var. *europaea* 'Sorani' | 0.02 | 0.03 | 0.37 | 0.06 | 0.07 | 0.39 | 0.38 |
| var. *sylvestris* | 0.05 | 0.05 | 0.37 | 0.06 | 0.07 | 0.40 | 0.38 |
| *maroccana* | 0.02 | 0.02 | 0.38 | 0.15 | 0.18 | 0.40 | 0.40 |
| *cerasiformis* | 0.05 | 0.05 | 0.36 | 0.15 | 0.18 | 0.40 | 0.39 |
| *guanchica* | 0.06 | 0.07 | 0.36 | 0.09 | 0.10 | 0.39 | 0.38 |
| *laperrinei* | 0.11 | 0.12 | 0.36 | 0.09 | 0.11 | 0.41 | 0.38 |
| *cuspidata* - R | 0.12 | 0.13 | 0.36 | 0.04 | 0.06 | 0.44 | 0.38 |
| *cuspidata* - S | 0.11 | 0.12 | 0.35 | 0.05 | 0.07 | 0.44 | 0.38 |
| *cuspidata* - I | 0.14 | 0.15 | 0.35 | 0.12 | 0.14 | 0.41 | 0.38 |

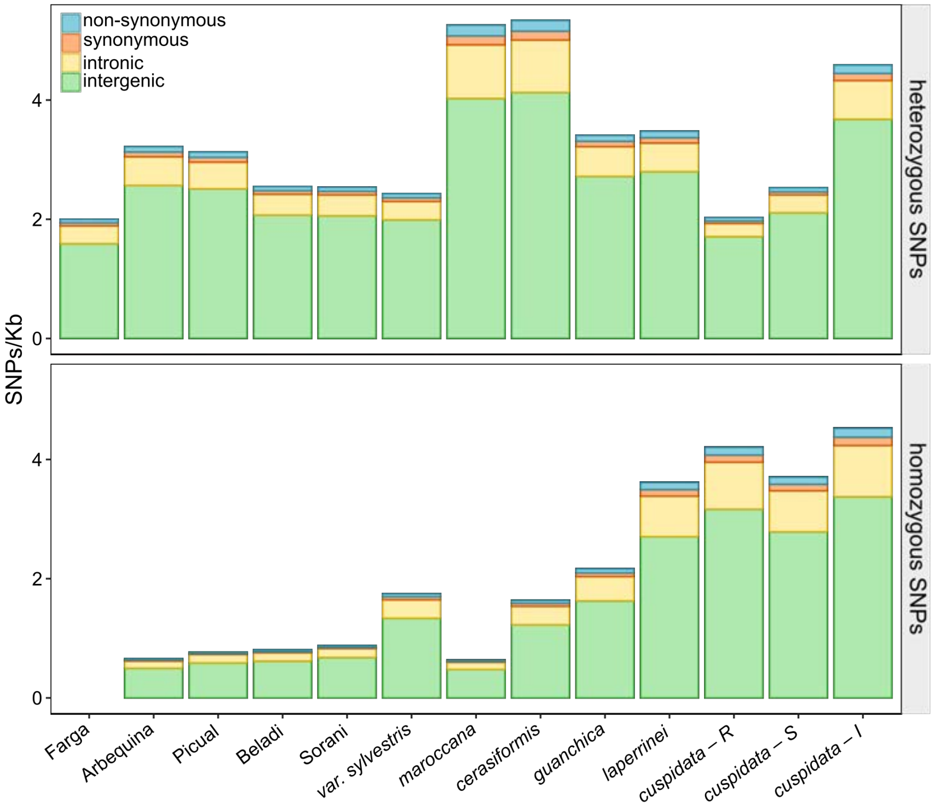

**Figure 4.2:** Number of homozygous and heterozygous SNPs (SNPs/Kb) in the intergenic, intronic and coding region of the genome. The coding region was divided according to the changes that the allele can produce (synonymous and nonsynonymous).

When we analyzed the SNPs that can produce a nonsynonymous change, including heterozygous and homozygous SNPs, we found that a total of 29,685 proteins (53% of the predicted proteome) have at least one SNP with nonsynonymous change, from which 9,704 are common for all the individuals (see Table 4.3). On the other hand, the list of proteins that did not show any nonsynonymous changes were used to do a functional enrichment analysis and we found two overrepresented GO terms: terpene synthase activity and defense response, indicating that these functional categories may be under a particularly strong purifying selection.

**Table 4.3:** Number of proteins with at least one nonsynonymous change per individual.

| *O. europaea* subsp. | Nº Proteins with Nonsyn SNPs |
|---|---|
| *europaea* var. *europaea* 'Farga' | 17,417 |
| *europaea* var. *europaea* 'Arbequina' | 20,664 |
| *europaea* var. *europaea* 'Picual' | 20,365 |
| *europaea* var. *europaea* 'Beladi' | 19,179 |
| *europaea* var. *europaea* 'Sorani' | 19,201 |
| *europaea* var. *sylvestris* | 20,676 |
| *maroccana* | 26,760 |
| *cerasiformis* | 25,333 |
| *guanchica* | 23,185 |
| *laperrinei* | 23,351 |
| *cuspidata* - R | 23,506 |
| *cuspidata* - S | 23,913 |
| *cuspidata* - I | 23,558 |

To further determine whether one or more proteins are under positive selection in cultivated olives, we used the set of proteins that have at least one nonsynonymous SNP, and we performed a selection test using the branch-sites model from codeml PAML package (see material and methods). For this test we used a pruned tree that included all the cultivars (var. *europaea*), the var. *sylvestris* and one subsp. *cuspidata* as the outgroup. We analyzed different hypotheses. A) five complementary models (A1 through A5) were assumed in which selection is acting in each cultivar, respectively; and B) there is selection in the common ancestor of all the cultivars (see Figure S4.4).

For the hypothesis of five complementary models (hypothesis A), we have a total of 60 proteins positively selected, with each cultivar having a specific set of selected proteins (ranging from 7 to 21, see Table 4.4 and Figure S4.5). Among these proteins, some have homologs associated with biotic and abiotic stress response (see Table 4.4). For example proteins associated with salt stress: in 'Picual' and 'Sorani' a U-box protein 30 (Hwang et al., 2014); and in 'Arbequina' and 'Beladi' a rhodanese-like domain-containing protein 4 (Wang et al., 2017a; Zhang et al., 2017a). In 'Farga', 'Arbequina', 'Picual', and 'Beladi' some ankyrins were also detected, which frequently take part in the defense response to pathogens (Vo et al., 2015; Mou et al., 2013). Specifically in 'Picual', two proteins were related with resistance to fungi, *R1* gene (Ballvora et al., 2002) and *Lr10 gene* (Feuillet et al., 2003). Interestingly, 'Picual' is susceptible to some fungi such as *Verticillium dahliae* (López-Escudero et al., 2004), but resistant to others such as *Colletotrichum acutatum* (Moral and Trapero, 2009; Cacciola et al., 2012). In 'Farga' a protein that regulates the activation of the immune response in *A. thaliana*, *MOS1* gene (Li et al., 2010; Zhu et al., 2010) was found to be under positive selection. Among other proteins, we also detected proteins positively selected that were related to metabolic and developmental processes. In 'Arbequina' and 'Picual' proteins positively selected were associated to auxin response factors (Ellis et al., 2005; Lim et al., 2010; Liu et al., 2015). In 'Picual' three proteins under selection were homologs to NIN-like protein 7 (NLP7) (Castaings et al., 2009; Karve et al., 2016), protein longifolia1 (Lee et al., 2006), 1-aminocyclopropane-1-carboxylate oxidase (ACO) (Ruduś et al., 2013). In 'Arbequina' two proteins were associated to caffeoyl-CoA O-methyltransferase (Wang et al., 2017c) and *Exo70A1* gene (Wang et al., 2013).

For the hypothesis of common ancestry of all cultivars (hypothesis B), we only have four proteins, which are associated with leucine-rich repeat receptor-like protein, auxin response factor, ankyrin repeat domain, and WEB family protein (Table 4.4). These four proteins are also present in the set of proteins under selection of some cultivars. Further work is needed, however, to better clarify the evolution and roles of these genes in the context of olive domestication.

**Table 4.4:** List of proteins under selection per each tested hypothesis and their associated function. Columns indicate, in the following order: the hypothesis, the olive protein name, the Id of the homologous protein, the species, and the description of the protein.

| Hypothesis | protein | homolog | species | description |
|---|---|---|---|---|
| A1 ('Farga') | OE6A015052 | XP_011081675.1 | *Sesamum indicum* | protein IQ-DOMAIN 1 |
| | OE6A018964 | XP_011101343.1 | *Sesamum indicum* | protein MODIFIER OF SNC1 1, partial |
| | OE6A025096 | EOY06149.1 | *Theobroma cacao* | RNA-binding family protein isoform 2 |
| | OE6A033957 | XP_011076334.1 | *Sesamum indicum* | probable pectinesterase/pectinesterase inhibitor 17 |
| | OE6A040753 | XP_011090054.1 | *Sesamum indicum* | monofunctional riboflavin biosynthesis protein RIBA 3, chloroplastic |
| | OE6A076031 | CDP21052.1 | *Erythranthe guttata* | GDSL esterase/lipase At1g29670-like |
| | OE6A078479 | XP_011099782.1 | *Sesamum indicum* | U-box domain-containing protein 28 |
| | OE6A098064 | XP_011095110.1 | *Sesamum indicum* | WEB family protein At2g38370 |
| | OE6A108230 | CDP05105.1 | *Coffea canephora* | unnamed protein product |
| | OE6A115491 | XP_011098830.1 | *Sesamum indicum* | ankyrin repeat domain-containing protein 50 |
| A2 ('Arbequina') | OE6A015333 | KNA24858.1 | *Erythranthe guttata* | plastidal glycolate/glycerate translocator 1, chloroplastic |
| | OE6A019061 | KGN62871.1 | *Cucumis sativus* | hypothetical protein Csa_2G378530 |
| | OE6A022634 | XP_016731875.1 | *Gossypium hirsutum* | cytochrome P450 CYP82D47-like |
| | OE6A022980 | XP_009407208.1 | *Musa acuminata* | caffeoyl-CoA O-methyltransferase |
| | OE6A025223 | XP_011089486.1 | *Sesamum indicum* | telomere repeat-binding protein 5 |
| | OE6A029165 | XP_015087654.1 | *Solanum pennellii* | ankyrin repeat domain-containing protein 13C-A-like |
| | OE6A031455 | XP_011091510.1 | *Sesamum indicum* | uncharacterized protein LOC105171936 |
| | OE6A038365 | XP_011074013.1 | *Sesamum indicum* | uncharacterized protein LOC105158828 |
| | OE6A041952 | CDP01914.1 | *Nicotiana tabacum* | zinc finger CCCH domain-containing protein 45-like |
| | OE6A044223 | XP_011072878.1 | *Sesamum indicum* | uncharacterized protein LOC105157993 |
| | OE6A047279 | XP_011082400.1 | *Sesamum indicum* | ABC transporter B family member 11-like |
| | OE6A050975 | XP_011076907.1 | *Sesamum indicum* | beta-D-glucosyl crocetin beta-1,6-glucosyltransferase-like |
| | OE6A054445 | XP_011080456.1 | *Sesamum indicum* | probable isoleucine–tRNA ligase, cytoplasmic |
| | OE6A067749 | XP_003633575.1 | *Vitis vinifera* | uncharacterized protein LOC100855398 |
| | OE6A069169 | CDP03993.1 | *Sesamum indicum* | probably inactive leucine-rich repeat receptor-like protein kinase At5g48380 |
| | OE6A086923 | XP_011093958.1 | *Sesamum indicum* | auxin response factor 18-like |
| | OE6A094656 | XP_011077281.1 | *Sesamum indicum* | exocyst complex component EXO70A1-like |
| | OE6A098047 | XP_011091404.1 | *Solanum pennellii*<br>protein 13C-A-like | ankyrin repeat domain-containing |
| | OE6A099135 | XP_011070370.1 | *Sesamum indicum* | oligopeptide transporter 7 |
| | OE6A106337 | XP_011076033.1 | *Sesamum indicum* | ethylene-responsive transcription factor CRF4-like |
| | OE6A121707 | XP_009761641.1 | *Nicotiana sylvestris* | rhodanese-like domain-containing protein 4, chloroplastic |
| A3 ('Picual') | OE6A019916 | XP_011097122.1 | *Sesamum indicum* | late blight resistance protein R1-A-like |
| | OE6A021997 | CDP18099.1 | *Erythranthe guttata* | glutamate receptor 1.3-like |
| | OE6A025223 | XP_011089486.1 | *Sesamum indicum* | telomere repeat-binding protein 5 |
| | OE6A025604 | XP_011073514.1 | *Sesamum indicum* | 1-aminocyclopropane-1-carboxylate oxidase homolog 1-like |
| | OE6A028717 | XP_011082828.1 | *Sesamum indicum* | leucine-rich repeat receptor-like serine/threonine-protein kinase At1g17230 |
| | OE6A031125 | CAA78386.1 | *Petunia x hybrida* | Protein 1 |
| | OE6A036294 | XP_011070353.1 | *Sesamum indicum* | protein NLP7 isoform X1 |
| | OE6A039384 | XP_011086137.1 | *Sesamum indicum* | lysine-specific demethylase JMJ25 |
| | OE6A039500 | XP_011076220.1 | *Sesamum indicum* | lipid phosphate phosphatase 2-like isoform X2 |
| | OE6A056603 | XP_011078077.1 | *Sesamum indicum* | uncharacterized protein LOC105161920 |
| | OE6A059201 | XP_017640072.1 | *Gossypium arboreum* | protein LONGIFOLIA 1-like |
| | OE6A076342 | XP_011093263.1 | *Sesamum indicum* | ankyrin repeat-containing protein At2g01680-like |
| | OE6A084635 | XP_011085170.1 | *Sesamum indicum* | U-box domain-containing protein 30-like |
| | OE6A093547 | XP_019158053.1 | *Ipomoea nil* | Leaf rust 10 disease-resistance locus receptor-like protein kinase-like 1.5 |
| | OE6A100401 | XP_011090586.1 | *Sesamum indicum* | protein EMBRYONIC FLOWER 1 isoform X1 |
| | OE6A110954 | KZV20136.1 | *Sesamum indicum* | alkaline/neutral invertase B |
| | OE6A114163 | XP_011073189.1 | *Sesamum indicum* | auxin response factor 2-like |

Continued on next page

**Table 4.4 – continued from previous page**

| Hypothesis | protein | homolog | species | description |
|---|---|---|---|---|
| A4 ('Beladi') | OE6A044961 | XP_011090510.1 | *Sesamum indicum* | protein UPSTREAM OF FLC |
| | OE6A046704 | XP_011081328.1 | *Sesamum indicum* | putative glycerol-3-phosphate transporter 4 isoform X1 |
| | OE6A049660 | XP_009780967.1 | *Nicotiana sylvestris* | peroxidase 41-like |
| | OE6A052800 | XP_016457035.1 | *Nicotiana tabacum* | probable LRR receptor-like serine/threonine-protein kinase At3g47570 isoform X1 |
| | OE6A052991 | CDP00074.1 | *Coffea canephora* | unnamed protein product |
| | OE6A098047 | XP_011091404.1 | *Solanum pennellii* | ankyrin repeat domain-containing protein 13C-A-like |
| | OE6A100346 | XP_011090426.1 | *Sesamum indicum* | zinc finger CCCH domain-containing protein 44-like isoform X1 |
| | OE6A113207 | XP_006356805.1 | *Solanum tuberosum* | 50S ribosomal protein 6, chloroplastic |
| | OE6A118614 | XP_011087574.1 | *Sesamum indicum* | protein NRT1/ PTR FAMILY 4.5-like |
| | OE6A121707 | XP_009761641.1 | *Nicotiana sylvestris* | rhodanese-like domain-containing protein 4, chloroplastic |
| A5 ('Sorani') | OE6A034615 | XP_011101657.1 | *Sesamum indicum* | glycosyltransferase family protein 64 protein C5-like |
| | OE6A054445 | XP_011080456.1 | *Sesamum indicum* | probable isoleucine–tRNA ligase, cytoplasmic |
| | OE6A057051 | XP_011076797.1 | *Sesamum indicum* | uncharacterized RING finger protein C4G3.12c-like |
| | OE6A066886 | XP_012840739.1 | *Erythranthe guttata* | triacylglycerol lipase 2 isoform X1 |
| | OE6A084635 | XP_011085170.1 | *Sesamum indicum* | U-box domain-containing protein 30-like |
| | OE6A098525 | XP_011081428.1 | *Sesamum indicum* | pre-mRNA-processing-splicing factor 8 |
| | OE6A100805 | XP_016552561.1 | *Capsicum annuum* | peroxisomal (S)-2-hydroxy-acid oxidase GLO1 |
| Hypothesis B | OE6A069169 | CDP03993.1 | *Sesamum indicum* | probably inactive leucine-rich repeat receptor-like protein kinase At5g48380 |
| | OE6A086923 | XP_011093958.1 | *Sesamum indicum* | auxin response factor 18-like |
| | OE6A098047 | XP_011091404.1 | *Solanum pennellii* | ankyrin repeat domain-containing protein 13C-A-like |
| | OE6A098064 | XP_011095110.1 | *Sesamum indicum* | WEB family protein At2g38370 |

### 4.3.3 Patterns of allelic representation in heterozygous positions suggest basal tetraploidy

In order to assess the ploidy of each individual we plotted the relative coverage of alternative alleles in heterozygous sites (see material and methods). In Figure 4.3 we can observe that the subsp. *maroccana* and *cerasiformis* show peaks consistent with their described hexaploid (peaks at 0.17, 0.33, 0.50, 0.67, 0.83) and tetraploid (peaks at 0.25, 0.50, 0.75) character, respectively (Besnard et al., 2008). Unexpectedly, the patterns for all the other sequenced genomes did not correspond to the expected single peak for a diploid (0.50). Instead, all supposed diploid subspecies showed patterns consistent with at least two peaks at frequencies around 0.25, and 0.50. The peak at 0.75, although very weak in the aggregate picture, was conspicuous when plotting some of the individual scaffolds (Figure S4.6). Furthermore, the relative density of the 0.25 and 0.50 peaks varied among the individuals.

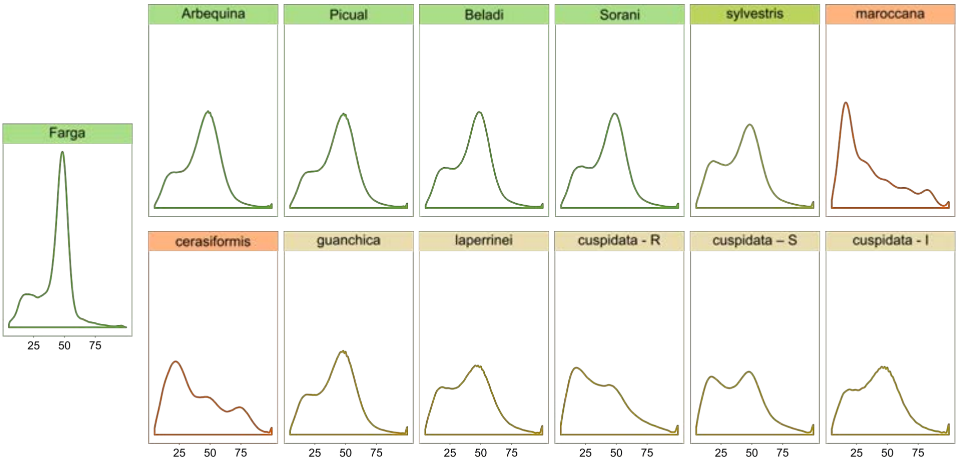

**Figure 4.3:** Density plot for the relative coverage of alternative alleles in heterozygous sites per each individual. The cultivated olives are marked in green, var. *sylvestris* in olive-green, *cerasiformis* and *maroccana* in orange, and the other subspecies in yellow. For all cases we only plotted data corresponding to the scaffolds larger than 100 Kb.

Only one peak (at 0.50) is expected for diploid regions of the genome, while peaks at 0.25 and 0.75 would only be present for tetraploid regions. Two scenarios may explain the existence of these regions. In one scenario, duplicated regions of the genome may have been collapsed. In another scenario, large regions of the genomes of the sequenced individuals are tetraploid. To investigate this further, we analyzed the K-mer spectra in the reference genome (cv. 'Farga') (Figure S4.7). The distribution of depth for each distinct k-mer shows a main homozygous peak, and a heterozygous peak at half the depth, as for the standard k-mer plots (e.g. Cruz et al. (2016a)). In this plot we also observe the amount of distinct K-mers absent (0x class, in black), as well as the copy classes present in the assembly. The absent elements are sequences that have not been assembled and therefore are not present in the reference. A good assembly will report a single haplotype, and thereby half of the bubbles in the heterozygous peak would be absent (e.g. Mapleson et al. (2017)). When the collapsing of alleles fails, this results in both haplotypes being present in the heterozygous part, and thus the homozygous k-mers around the bubble are duplicated in the assembly. Therefore, uncollapsed alleles end up as artefactual duplications that are present twice and have a similar depth to the homozygous 27-

mers (2x class, see the violet areas above the peak of higher depth). The violet peak of artefactual duplications below the homozygous peak (depth 31 to 74 approximately) represents approximately 2.78% of the assembly and corresponds with heterozygous 27-mers that have been uncollapsed and therefore artificially duplicated (Figure S4.7). The left-most part of this distribution (depth <31 and >7) could be interpreted as homozygous tetraploid regions that have been unfolded (4n-HET), while above depth 74 the sequence is likely to be triploid (3n). There is a small amount of sequence that appears 4x times in the genome and has the same coverage as the homozygous 27-mer peak (from depth 31 to 74). This can be interpreted as tetraploid sequences that have been collapsed in the assembly (4n-HOM). These results indicate that the reference genome have few regions collapsed.

In order determine if the regions with signs of tetraploidy are product of collapse in the reference genome we analyzed the coverage of the positions with frequencies 0.25, 0.5 and 0.75 in all the individuals (see material and methods), and we observed a different pattern: for the tetraploid individual (*cerasiformis*), all the frequencies have a similar median coverage, while for all the other individuals the positions with frequencies at 0.50 have lower coverage than the positions at 0.25 and 0.75 (Figure S4.8), although they overlapped significantly. Despite the fact that higher coverage in positions with alleles at 0.25 and 0.75 frequencies is expected in the case of artefactual collapsing of duplicated regions, this difference in coverage is far from being double as it would be expected for collapsed regions of the genome. We next analyzed patterns in contigs assembled from pools of fosmid libraries. In these contigs there is a high chance of capturing a single haplotype of the source genome. As we can see in Figure S4.9, we found regions with peaks between 0.25, 0.5, and 0.75 in these contigs. Moreover, although the relative density of each of the peaks varied from scaffold to scaffold, most of the analyzed scaffolds showed such patterns suggestive of tetraploidy. This indicates that the affected regions are widespread and affect a major fraction of the assembly. Three whole genome duplications have been described in the lineage leading to *O. europaea* (Julca et al., 2017), of which one is fairly recent, having occurred after the divergence of *Olea* from *Phillyrea angustifolia* ($\sim$10 Mya). Genes duplicated in a recent polyploidization are candidates for having been collapsed during the assembly process. If that would be the case, then peaks at 0.25 and 0.75 should not be observed in detected duplicates as they constitute different loci and are therefore uncollapsed (otherwise they would have been detected as a single gene). As seen in Figure S4.10, recent duplicates also contained these peaks suggesting that the collapsing of regions from this polyploidization is not the cause of the unorthodox peaks. Considering all these data we conclude that the level of ploidy is higher than two for the most part of the genome in all analyzed individuals. Furthermore, although we cannot discard the possibility that some genomic regions have been collapsed in our reference assembly, we attribute the major part of the signal to the existence of a tetraploid state.

Interestingly the peaks at 0.25 in duplicated genes are more evident than in non-duplicated genes (Figure S4.10) suggesting that the non-duplicated genes have a tendency to show evidence of two chromosome complements (diploidy). Some studies have shown that some genes in angiosperms after polyploidizations tend to be preserved as singletons, generally calling them as "duplication-resistant" genes (Paterson et al., 2006; De Smet et al., 2013). This kind of phenomena may explain the observed plots in the non-duplicated genes.

Importantly, the genomic pattern of tetraploidy is observed in all the historically considered "diploid" lineages suggesting that this polyploidization should have taken place before the divergence of the sequenced subspecies. This assumption is more parsimonious than assuming many independent events. If this is true, the two polyploid individuals, *cerasiformis* and *maroccana*, are probably the result of more recent events. The subsp. *maroccana* could be originated by an hybridization between a diploid x tetraploid, or two triploid individuals. However, the origin of *cerasiformis* is much more difficult to explain. One possibility is that *cerasiformis* is the result of a very recent polyploidy, where the polymorphic positions did not have enough time to diverge.

### 4.3.4 Phylogenetic relationships in the *O. europaea* complex

In order to understand the phylogenetic relationships within the *O. europaea* complex, we ran a model-based genetic structure analysis using nuclear SNPs, and reconstructed phylogenetic trees using nuclear, plastid and mitochondrial SNPs, separately (see material and methods). Structure analysis results showed that the most likely number of genetic groups (k) among the *O. europaea* complex is k = 3 (Figure 4.4). These three clusters of genetic ancestry are differentially present among the sequenced individuals (Figure 4, Table S4.3). Only one of the groups (3) is exclusively present in one of the subspecies (*cuspidata*). One of the genetic clusters (1) is primarily found in the var. *sylvestris*, particularly among cultivars (var. *europaea*), but it is also present to a lesser proportion in the other subspecies (*cerasiformis*, *guanchica*, *laperrinei*, *maroccana*), and in the *cuspidata* individual from Iran. The remaining genetic cluster (2) is more abundant in *cerasiformis*, *guanchica*, *laperrinei* and *marocana*, but forms a significant fraction of the genetic background of the wild individual of the subspecies *europaea* and the sample of *cuspidata* from Iran. This latter individual shows a similar proportion of the three identified genetic clusters, which contrasts with the other *cuspidata* samples that are almost purely presenting cluster 3. Such mixed ancestries inferred from genetic data can be interpreted as arising from recent admixture among multiple founder populations, but they can also be the result of shared ancestry before the divergence of the populations (Li et al. (2008); but see Mousavi et al. (2017)).

The dominance of one of the genetic backgrounds among cultivars could also result from domestication, since the admixture proportion of this genetic cluster (1) increases from the subspecies *laperrinei*, *maroccana*, *cerasiformis* and *guanchica* to the oleaster, becoming dominant in the cultivated olives (Figure 4 and Table S4.3). This effect could have been achieved from preferential selection of genetic variants among the standing variation or from selective crosses with a population with such background. The differences of the individuals of subsp. *cuspidata* are also remarkable. Two individuals from the Reunion island and South Africa (*cuspidata*-R and *cuspidata*-S) belong to a single genetic group, while the *cuspidata* tree from Iran (*cuspidata*-I) show similar proportions of three genetic groups. As we expect the *cuspidata*-R forms a unique cluster (Q = 100%) since it comes from an island and could for instance have limited genetic flux. The *cuspidata*-S shares the same genetic group of *cuspidata*-R (Q = 96%), with a small part of genetic group 2, which reflects limited contact with other subspecies of the *O. europaea* complex. However, the *cuspidata*-I is a mixture of the three clusters (Figure 4.4), which indicates that there were multiple contacts with other lineages from Africa and Europe. Differences between *cuspidata* from Africa and Asia were observed in previous studies based on organellar markers (Besnard et al., 2002a,b, 2001b; Lumaret et al., 2000; Besnard et al., 2007b). These studies have shown that two geographically distant groups exhibit two different chlorotypes: "A" in Tropical and Southern Africa, and "C" in Southern Asia to Eastern Africa. Furthermore in some regions of Iran, *cuspidata* populations occur close to cultivated olive suggesting the possibility of sporadic hybridization among them (Sheidai et al., 2010; Hosseini-Mazinani et al., 2014; Besnard and El Bakkali, 2014).

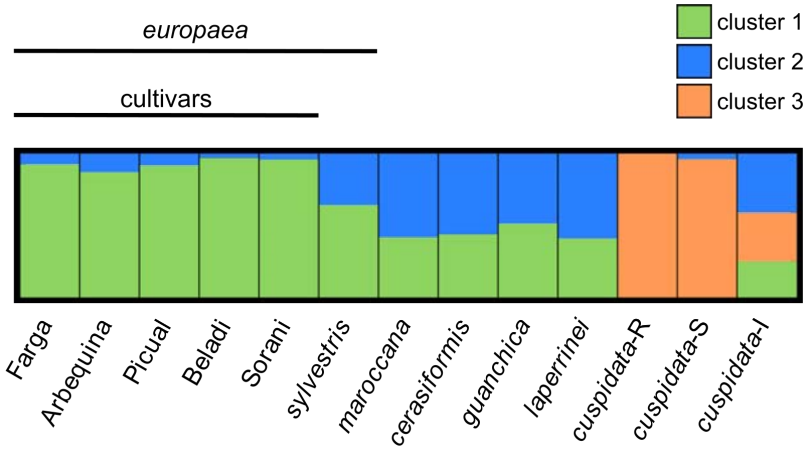

**Figure 4.4:** Bayesian clustering for the SNP data estimated in Structure v2.3 for the *O. europaea* complex. Structure bar plot showing three genetic clusters differentiated by colour.

Phylogenetic analyses, and previously proposed phylogenetic relationships, suggest the existence of rampant genetic flux in the evolution of the *O. europaea* complex. The split network tree (Figure 4.5) reveals a heavily reticulated structure with conflicting phylogenetic signals affecting mostly the relationships among cultivars and among wild subspecies other than *cuspidata*. In this network we notice that the most differentiated group is subsp. *cuspidata*, which is compatible with the proposed earliest divergence of this group (Besnard et al., 2011; Rubio de Casas et al., 2006; Besnard et al., 2009). We also reconstructed phylogenies from whole genomic information, which are supposed to represent the dominant phylogenetic signal, using nuclear, plastid, and mitochondrial SNPs, separately (Figure 4.6). In the plastid tree (Figure 4.6b) we can observe that the individuals group according to the previous described chlorotypes (Besnard et al., 2011, 2007b). Moreover the topology of the phylogenetic tree does not change if we use other plastid reference genomes for the SNP calling (data not shown). However, between the organellar and nuclear trees we can observe some incongruences (Figure 4.6), which is provided in other studies as evidence of hybridization processes (Fehrer et al., 2007; Barber et al., 2007; Pelser et al., 2010).

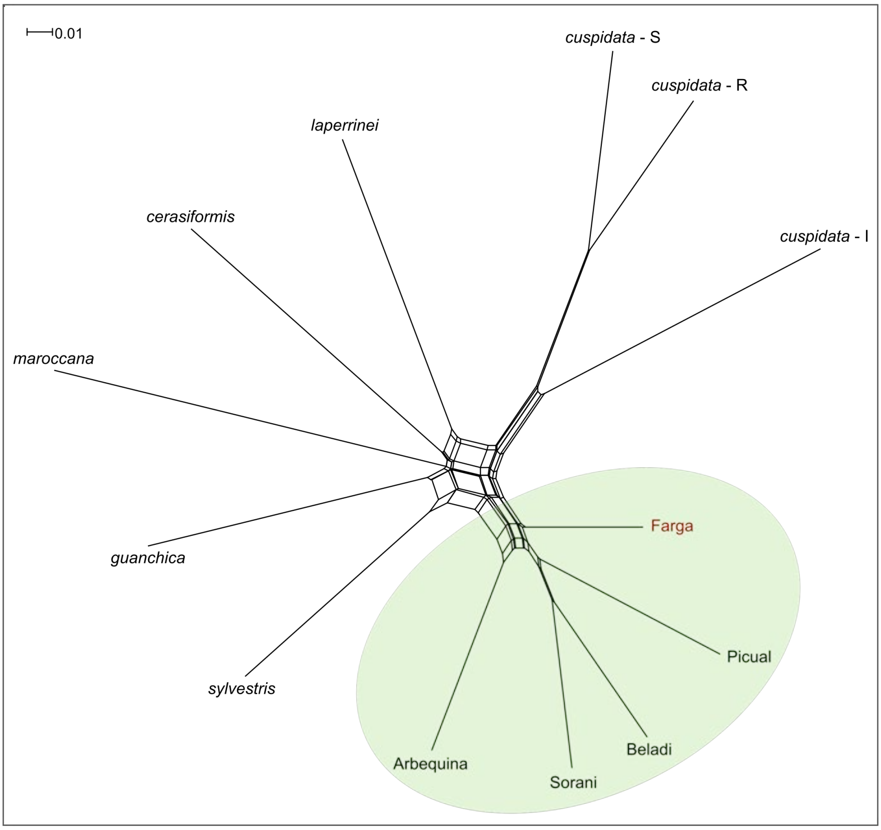

**Figure 4.5:** SplitsTree derived of nuclear SNPs. All the cultivars are marked in green, and the reference genome in red. The neighbor-net method is used here to explore data conflict and not to estimate phylogeny.

One incongruence observed involves the subsp. *laperrinei*, which is sister to *cerasiformis* in the nuclear tree (Figure 4.6a), but this subspecies is found closer to the cultivars that have the eastern Mediterranean chlorotype (E1) in both organellar trees (Figure 4.6b,c). This is in agreement with previous studies based on nuclear markers and plastid genomes, which reveal historical hybridization between *laperrinei* and *europaea* (Besnard et al., 2001b, 2002b, 2007b, 2013a; Besnard and Bervillé, 2000; Angiolillo et al., 1999; Rubio de Casas et al., 2006). This incongruence was better explained by hybridization between these two subspecies during waning and waxing of African lineages following climatic fluctuations during the Pleistocene (Besnard et al., 2009, 2007b; Rubio de Casas et al., 2006).

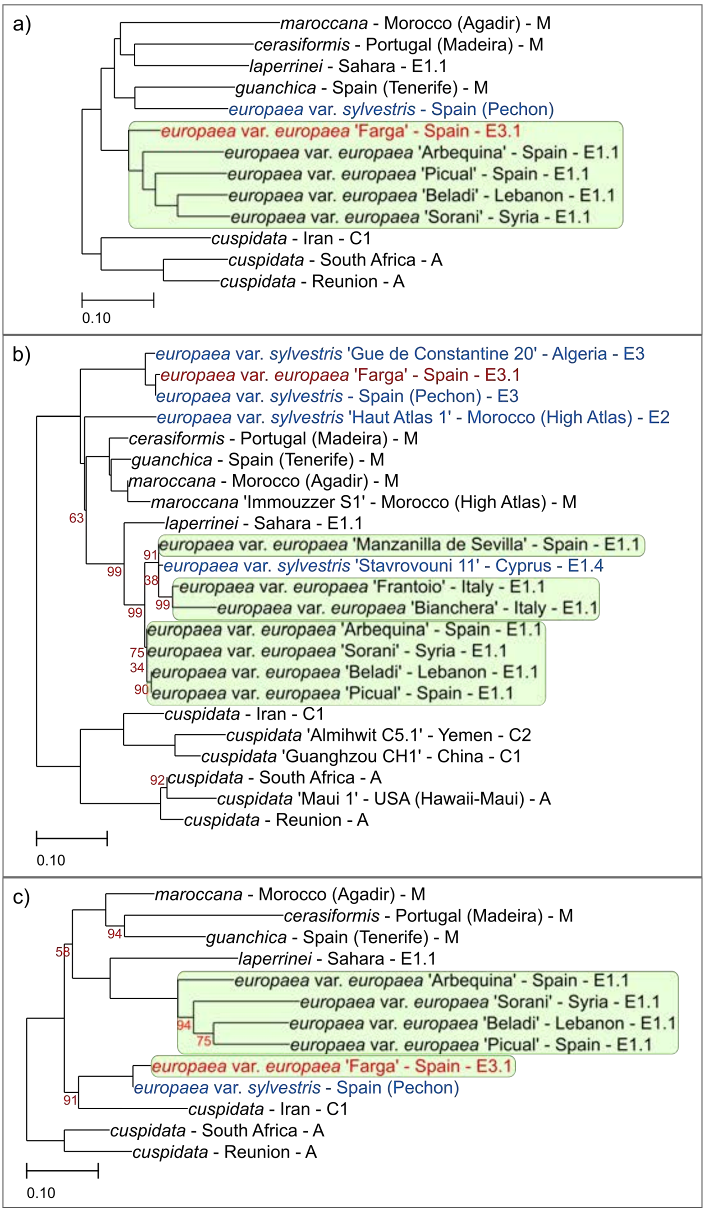
a)
maroccana - Morocco (Agadir) - M
cerasiformis - Portugal (Madeira) - M
laperrinei - Sahara - E1.1
guanchica - Spain (Tenerife) - M
europaea var. sylvestris - Spain (Pechon)
europaea var. europaea 'Farga' - Spain - E3.1
europaea var. europaea 'Arbequina' - Spain - E1.1
europaea var. europaea 'Picual' - Spain - E1.1
europaea var. europaea 'Beladi' - Lebanon - E1.1
europaea var. europaea 'Sorani' - Syria - E1.1
cuspidata - Iran - C1
cuspidata - South Africa - A
cuspidata - Reunion - A
0.10
b)
europaea var. sylvestris 'Gue de Constantine 20' - Algeria - E3
europaea var. europaea 'Farga' - Spain - E3.1
europaea var. sylvestris - Spain (Pechon) - E3
europaea var. sylvestris 'Haut Atlas 1' - Morocco (High Atlas) - E2
cerasiformis - Portugal (Madeira) - M
guanchica - Spain (Tenerife) - M
maroccana - Morocco (Agadir) - M
maroccana 'Immouzzer S1' - Morocco (High Atlas) - M
63
laperrinei - Sahara - E1.1
europaea var. europaea 'Manzanilla de Sevilla' - Spain - E1.1
91
europaea var. sylvestris 'Stavrovouni 11' - Cyprus - E1.4
99
38
europaea var. europaea 'Frantoio' - Italy - E1.1
99
europaea var. europaea 'Bianchera' - Italy - E1.1
99
europaea var. europaea 'Arbequina' - Spain - E1.1
75
europaea var. europaea 'Sorani' - Syria - E1.1
34
europaea var. europaea 'Beladi' - Lebanon - E1.1
90
europaea var. europaea 'Picual' - Spain - E1.1
cuspidata - Iran - C1
cuspidata 'Almihwit C5.1' - Yemen - C2
cuspidata 'Guanghzou CH1' - China - C1
92
cuspidata - South Africa - A
cuspidata 'Maui 1' - USA (Hawaii-Maui) - A
cuspidata - Reunion - A
0.10
c)
maroccana - Morocco (Agadir) - M
cerasiformis - Portugal (Madeira) - M
94
guanchica - Spain (Tenerife) - M
58
laperrinei - Sahara - E1.1
europaea var. europaea 'Arbequina' - Spain - E1.1
europaea var. europaea 'Sorani' - Syria - E1.1
94
europaea var. europaea 'Beladi' - Lebanon - E1.1
75
europaea var. europaea 'Picual' - Spain - E1.1
europaea var. europaea 'Farga' - Spain - E3.1
europaea var. sylvestris - Spain (Pechon)
91
cuspidata - Iran - C1
cuspidata - South Africa - A
cuspidata - Reunion - A
0.10

**Figure 4.6:** Maximum likelihood species tree derived from the SNPs data: a) nucleus with 13,543,130 positions, b) plastid with 319 positions, and c) mitochondria with 2,614 positions. In red the reference genome for the SNP calling. In green all the individuals belonging to the var. *europaea* (cultivars), and in blue all the var. *sylvestris*. The geographical location of the sample and the chlorotype are indicated. All bootstrap values that are not maximal (100%) are indicated in the figure.

Other phylogenetic incongruence was observed in cv. 'Farga', which clusters together with the other cultivated olives in the nuclear tree, but is sister of the var. *sylvestris* from Spain (Pechón) and thus far from the other cultivars in both plastid and mitochondrial trees. This result supports previous evaluation of maternal inheritance of plastid and mitochondria (Besnard et al., 2000). In addition, these results suggest that the maternal line of 'Farga' originated from wild olive trees from the western Mediterranean basin (which carry the E3 chlorotype), and the paternal line from previous domesticated individuals from the eastern Mediterranean basin. In a previous study that combined a large sample of cultivated olives and oleasters, a similar pattern was observed, in which most cultivars were assigned to the eastern genetic pool, even those with maternal lineages that originated from the western Mediterranean basin (Besnard et al., 2013a,b). All these results reinforce the idea that cultivars are either from the eastern genetic pool or admixed forms (Besnard et al., 2013b; Besnard and Rubio de Casas, 2016; Kaniewski et al., 2012), and support secondary domestication process in the western Mediterranean basin.

Not all the accessions of subsp. *europaea* cluster together. In the nuclear tree (Figure 4.6a) we can see that all the cultivars group together, while the var. *sylvestris* clusters separately and closer to the subsp. *guanchica*. In the plastid and in the mitochondrial trees we can observe that the cv. 'Farga' and the var. *sylvestris* from Pechón (Spain) cluster together (both have a E3 western like chlorotype), but far from the other cultivars holding the eastern like chlorotype (E1). These results suggest that the subsp. *europaea* is polyphyletic. Some other studies with plastid and nuclear markers show similar results with regard to this subspecies (Besnard et al., 2009, 2007b; Vargas and Kadereit, 2001).

In the nuclear and the plastid trees, the three individuals of subspecies *cuspidata* form a monophyletic group (Figure 4.6a,b). Interestingly, they are divided in two branches: *cuspidata*-S and *cuspidata*-R on one side, and *cuspidata*-I on the other. In the mitochondrial tree (Figure 4.6c), *cuspidata*-I shows a different pattern and is sister group of the subsp. *europaea*, which is inconsistent with the plastid tree. This incongruence should be considered carefully because can be the result of a nuclear contamination in the mitochondrial genome assembly. It is true that in our analysis we did not include many individuals of the subsp. *cuspidata*, but from our results we can see that the individuals from Africa are highly differentiated from the individual from Iran. Furthermore the individual from Iran seems to be the result of admixture between the three detected genetic clusters as was explained before (Figure 4.4). Interestingly a recent analysis of plant reproductive structures in Asian and African *cuspidata* accessions has shown numerous differences at morpho-structural and functional levels, but this variability was suggested to be due to a different adaptability to the growth environment (Caceres et al., 2016). Considering the observed large genetic differences between the Iran individual of *cuspidata* and the other *cuspidata* accessions, comparable to those existing between other established subspecies, the subdividision of *cuspidata* into more than one taxon should be considered. In summary, our results show strong patterns of hybridization among all the lineages of the *O. europaea* complex and corroborate the idea that the main force for lineage divergence is geographical isolation (Besnard et al., 2002b; Rubio de Casas et al., 2006; Besnard et al., 2009; Besnard and El Bakkali, 2014). However, the addition of more individuals is needed for any taxonomic consideration.

Phylogenetic analyses also indicate that the polyploid subspecies *cerasiformis* (4x) and *maroccana* (6x) of the *O. europaea* complex are closely related to subspp. *laperrinei* (2x) and *guanchica* (2x) in the three-genome phylogenies (Figure 4.6). In the nuclear tree the two polyploid subspecies form a monophyletic group with the subsp. *laperrinei*, while in the organellar trees they cluster with the subsp. *guanchica*. Our results are in agreement with previous plastid phylogenies (Besnard et al., 2002b), but not with a nuclear phylogeny based on ITS pseudogene sequences (Besnard et al., 2009). Moreover, previous studies have found that a single subspecies (*laperrinei*) have two ploidy levels (2x, 3x) (Besnard and Baali-Cherif, 2009). These phylogenetic results show evidence that hybridizations with change in ploidy level from an ancestor shared with subsp. *laperrinei* and *guanchica* may have brought about the two polyploid subspecies. In any case, these four subspecies are involved in a ploidy series (2x, 4x, 6x) suggesting that contemporary polyploidization may reflect similar polyploidization processes millions of years ago.

### 4.3.5 Concluding remarks

This study presents the first phylogenomic analysis of the *O. europaea* complex. Our main results show that the cultivated individuals have lower nucleotide diversity compared with the wild individuals (oleaster and other subspecies), although this is still quite high for a cultivated plant. Some genes positively selected in cultivated olives are associated with a response to biotic and abiotic stress and developmental processes, probably as product of domestication. The relative coverage of alternative alleles in heterozygous sites analysis provides evidence for a recent polyploidization in *O. europaea*, preceding the divergence of the subspecies. In addition, admixture and phylogenies analysis show that hybridization processes shaped the evolutionary history of the different lineages of the *O. europaea* complex. Particularly, the cv. 'Farga' has a different phylogenetic origin than the other cultivars suggesting a secondary domestication event in the Spanish area, in which var. *sylvestris* from Pechón acted as the maternal line, while a previous domesticated olive, as the paternal line. However, an increased sampling is needed to help describe general patterns of hybridization in *Olea europaea* and its evolution across the numerous areas of the Mediterranean basin.

## 4.4 Material and Methods

### 4.4.1 Genome sequences

We sampled twelve trees belonging to the six defined subspecies of the *Olea europaea* complex and from different locations, thus covering not only the taxonomy but also the geography and diversity of the species. Our sampling includes four individuals of var. *europaea* (cv. 'Arbequina', cv. 'Beladi', cv. 'Picual' and cv. 'Sorani'), one of var. *sylvestris*, one of *cerasiformis*, three of *cuspidata*, one of *guanchica*, one of *laperrinei*, and one of *maroccana* (see Table S4.1). The DNA of these twelve individuals was extracted as described in Cruz et al. (2016a) and their genomes were fully sequenced using Illumina HiSeq 2000 pair-end technology to a sequencing depth ranging from 24 to

34x at the CNAG-CRG sequencing facilities, as described for the reference genome (Cruz et al., 2016a). In addition to these twelve individuals, we used public data of the reference genome sequence of *Olea europaea* (cv. 'Farga') (Cruz et al., 2016a), and downloaded ten *O. europaea* plastid genomes from the NCBI database (see Table S4.1).

### 4.4.2 Assembly of the plastid and mitochondrial genomes

The available reference genome sequence does not include separate scaffolds for mitochondrial and plastid genomes (Cruz et al., 2016a). Here, we assembled and annotated both organellar genomes of the cv. 'Farga' using paired-end data from the reference genome sequence project (Cruz et al., 2016a). Briefly, for the plastid genome we mapped all whole genome shotgun (WGS) illumina reads (760bp insert lib) to the reference chloroplast sequence (NC_013707) and selected those pairs where at least one mate read maps (up to 4% mismatch). Subsequently, we assembled mapped reads with ABySS v1.3.6 (Simpson et al., 2009) (k=97) and obtained a fragmented assembly totalling 142 kb of sequence. Finally we used RAGOUT v2.0 (Kolmogorov et al., 2014) to produce an assisted assembly (i.e. ordering the contigs according to the reference). For the mitochondrial genome, we mapped all WGS illumina reads (760bp insert library) to mitochondrial genomes of two related species (Species–NCBI accessions: *Hesperelaea palmeri*–KX545367, and *Mimulus guttatus*–JN098455) and filtered the pairs as described above. Then we used SPAdes v.3.1.1 (Bankevich et al., 2012) for the assembling and SSPACE-LongRead (Boetzer and Pirovano, 2014) for the scaffolding steps. Finally, gaps were filled with gapcloser (Luo et al., 2014). For the plastid genome, the final assembly has a size of 136,336 base pairs (bp), which is smaller than the previous olive plastid genomes sequenced, which range from 155,531 bp to 155,896 bp (Besnard et al., 2011; Mariotti et al., 2010). The 20 kb of missing sequence correspond to a nearly identical inverted repeat. For the mitochondrion we recovered 593,378 bp divided in three scaffolds. This represents the first partial assembly of a mitochondrial genome in *O. europaea*.

The partial plastid genome was annotated using DOGMA (Dual Organellar GenoMe Annotator) (Wyman et al., 2004), which identifies putative genes by performing BLAST searches against a custom database. The start and stop codons, as well as the intron and exon boundaries were selected manually and based on a BLAST search against three available annotated plastid genomes of *O. europaea* (cultivar–NCBI accessions: 'Bianchera'–NC_013707, 'Frantoio'–GU931818, 'Manzanilla'–FN996972). The annotation of the partial mitochondrial genome was done by BLAST searches using twelve annotated mitochondrial genomes (NCBI accessions: KF709392, Y08501, JN107812, NC_029182, KT959112, KF815390, KX545367, JN098455, BA000042, KF177345, KY774314, KC189947) and the gene structures (i.e. intron-exon boundaries) were defined using Exonerate v2.47.3 with the "protein2genome" model (Slater and Birney, 2005). The annotations of tRNA genes were verified using tRNAscan-SE (Lowe and Chan, 2016).

In the plastid genome, we annotated 120 genes out of the 130–133 genes reported in other olive plastid genomes (Figure S4.3) (Besnard et al., 2011; Mariotti et al., 2010). From these genes, 80 are protein coding genes, 33 transfer RNA, and 7 ribosomal RNA (see Table S4.2 Figure S4.11). The coding regions in the olive mitochondrion comprise 40 protein-coding genes, 3 ribosomal RNA genes, and 12 transfer RNA genes (Figure S4.11). These genes represent approximately 80% of the genes annotated in *Hesperelaea palmeri* (68 genes) (Van de Paer et al., 2016) and *Mimulus guttatus* (62 genes) (Mower et al., 2012).

### 4.4.3 Detection of Single Nucleotide Polymorphisms

In order to compare the nucleotide diversity across sequences of the *O. europaea* complex at nuclear, plastid and mitochondrial level, we called SNPs using the cv. 'Farga' genome as a reference. Sequenced reads from each individual were mapped against the respective reference genome using BWA 0.7.6a-r433 (Li and Durbin, 2009), and SNPs were identified with GATK HaplotypeCaller v3.5 (McKenna et al., 2010), setting ploidy according to the described ploidy level of the individual, i.e. hexaploid for *maroccana*, tetraploid for *cerasiformis*, and diploid for the other four subspecies; and using thresholds for mapping quality (>40) and read depth of coverage (>20). Given the high number of the nuclear SNPs, a vcf file using SAMtools mpileup was created (Li et al., 2009). We applied an additional filter, which required that each polymorphic position should pass the read depth of coverage filter (>20) in all the individuals. SNPs calling was also explored changing the ploidy level of diploids into tetraploids. The results were consistent in 95% of the positions and thus only results obtained using ploidy level 2 are shown. To assess ploidy in single haplotype sequences we also called SNPs in contigs assembled from fosmid pool libraries following the same parameters as described before. Only contigs larger than 10 Kb were included for the fosmid assembly.

### 4.4.4 SNPs characterization

Nei's gene diversity index (Nei, 1973) of nuclear, plastid and mitochondrial SNP data was used to estimate nucleotide diversity. We kept the alternative allele in all the cases of nuclear heterozygous positions because we did not have a phased genome assembly.

The nuclear SNPs were classified according to their position in the genome as intergenic, intronic and coding. Coding SNPs were further classified into synonymous and nonsynonymous, according to the implied change in the respective codon. For the heterozygous positions, if at least one of the changes was nonsynonymous we considered the position as nonsynonymous. GO term enrichment analyses of the proteins without nonsynonymous SNPs was calculated using FatiGO (Al-Shahrour et al., 2004).

To investigate the variation of nonsynonymous and synonymous SNPs in the coding regions, we compared nonsynonymous changes per nonsynonymous site ($\pi$N) to synonymous changes per synonymous site ($\pi$S) by assuming that 75% of all sites are nonsynonymous.

### 4.4.5 Selection Tests

In order to detect the protein-coding genes that have potentially undergone selection among the cultivated individuals we used a subset of individuals, including all the cultivars ('Farga', 'Arbequina', 'Beladi', 'Sorani', 'Picual'), the var. *sylvestris*, and a subsp. *cuspidata* sample as the outgroup. Branch-site model implemented in the codeml PAML package v4.9 (Xu and Yang, 2013) was tested in all the proteins that had at least one nonsynonymous position. We marked the branches according to two hypotheses (see Figure S4.4): A) each terminal branch leading to of each cultivar was marked independently, B) the branch subtending the common ancestor of all the cultivars was marked. In addition, a multiple test analysis was performed using the multtest package from R (Pollard et al., 2007). Finally, we took all candidate proteins for positive selection and performed BLAST searches against the NCBI non-redundant protein sequences (NR) database (Camacho et al., 2009).

### 4.4.6 Analysis of K-mer Spectra in the Reference Genome Sequence

In order to evaluate the level of artefactual duplications present in the cv. 'Farga' assembly (Oe6) we used the Kmer Analysis Toolkit (Mapleson et al., 2017). This program was used to obtain a stacked histogram based on the 27-mer matrix of the assembled genome and the PE725 library (Cruz et al., 2016a). This library was selected because it has sufficient coverage (96.3x) and is evenly distributed across the genome. These plots are typically used to compare a Jellyfish hash (e.g. Marçais and Kingsford (2011)) produced from a read set, to a Jellyfish hash produced from an assembly. We plotted a decomposition of the stacked histogram for 'Farga' assembly version Oe6 for clarity (Figure S4.7).

### 4.4.7 Ploidy estimation

To assess the ploidy of each sequenced individual we used the nuclear SNP data and plotted the relative coverage of alternative alleles in heterozygous sites. We considered only scaffolds longer than 100 Kb. We used three alternative methods to obtain and filter heterozygous SNPs: A) directly from the vcf files obtained with GATK, B) for each heterozygous positions obtained with GATK, we computed allelic depths from the vcf file created with SAMtools mpileup, C) we applied an extra filter to the vcf file created with SAMtools mpileup, and for the heterozygous positions we included only the positions where the alternative alleles have at least 10% of the total depth. In the three cases the relative coverage of alternative alleles was obtained by dividing the alternative allelic depth by the total depth at that position. For a diploid genome, we would expect a single peak around 0.50 at biallelic positions, for a tetraploid three peaks, around 0.25, 0.50, and 0.75, and for an hexaploid five peaks around 0.17, 0.33, 0.5, 0.67, 0.83. The three methods gave consistent results and therefore, only results obtained by the second method are shown.

For individuals with peaks around 0.25, 0.50, and 0.75 we selected the positions around these frequencies ( 0.05 range) and plotted box plots of the total read depth after excluding positions with depths >200. For all the plots we used the ggplot2 package from R (Wickham, 2009).

In order to assess whether the obtained ploidy estimation is an assembly artefact caused by collapsed duplicated regions, we decided to compare the ploidy in coding regions. For this, we computed the allelic frequencies for all the duplicated genes versus all the non duplicated genes, assuming that the recently duplicated genes that are detected in the annotation should be in uncollapsed regions and should show a diploid pattern (as otherwise they could have not been annotated as different loci). The list of recently duplicated genes (age 1, corresponding to olive specific duplications) was obtained from the phylome analysis described in our previous study (Julca et al., 2017), and available in PhylomeDB (Huerta-Cepas et al., 2014).

### 4.4.8 Admixture Mapping

Because of the large number of polymorphic positions in the nuclear genomes of the *O. europaea* complex, it was only computationally feasible to analyze 100,000 positions. We generated 10 subsets of 100,000 randomly-chosen polymorphic positions without overlaps, and analyzed them in parallel. Then we identified population structure without a priori grouping assumptions, using the Structure software v2.3.4 (Pritchard et al., 2000). Structure was run with 100,000 generations of 'burn in' and 100,000 Markov chain Monte Carlo (MCMC) iterations after burn-in for increasing K values ranging from 1 to 7, considering independent alleles and admixture of individuals. Simulations were repeated 10 times for each value of K. The optimal number of genetic clusters was determined using $\Delta K$

method (Evanno et al., 2005) with the software Structure Harvester (Earl and VonHoldt, 2012). Finally, the optimal K value was visualized with DISTRUCT v1.1 (Rosenberg, 2004).

### 4.4.9 Phylogenetic analysis

Phylogenetic trees were reconstructed using SNPs data from nuclear, plastid and mitochondrial genomes, separately. In all cases, the genome sequence of the sequenced individuals was obtained by replacing the SNP positions in the respective reference genome, resulting in a pseudoalignment of all the considered genomes. Specifically, for the heterozygous SNPs of the nuclear dataset we randomly selected one allele per each position. For the plastid genomes, we included additional sequences by aligning our genomes with the genomes available in the databases (see Table S4.1) using MAFFT v7.305b (Katoh et al., 2005). All these alignments were trimmed using trimAl v1.4 (Capella-Gutiérrez et al., 2009) with the options -st 1 and -complementary, in order to remove all the non-informative positions. The final alignment had 13,543,130 variable positions for the nuclear genome, 319 for the plastid genome, and 2,614 for the mitochondrial genome. Phylogenetic trees were reconstructed from these alignments using PhyML v3.1 (Guindon and Gascuel, 2003) and the GTR model because it has been demonstrated as the most frequent evolution model in angiosperms. Support values based on 100 bootstraps replicates were calculated. A phylogenetic network using SplitsTree4 v4.14.5 and the NeighborNet approach was also reconstructed for the nuclear data (Huson and Bryant, 2006).

### Acknowledgements

We want to acknowledge Guillaume Besnard, Manuel Nogales, Concepción Muñoz Díez, and Carlos Garcia-Verdugo for providing some of the plant samples used in this study. TG and PV acknowledge support from Banco Santander for the olive genome sequencing project. IJ was supported in part by a grant from the Peruvian Ministry of Education: 'Beca Presidente de la República' (2013-III).

## 4.A Supplementary material

**Table S4.1:** *O. europaea* genomes used in the analysis. The columns show the sample origin, haplotype, ploidy level, and the source of the data.

| *O. europaea* subsp. | Origin | Chlorotype | Ploidy level | Source |
|---|---|---|---|---|
| var. *europaea* cv. 'Farga' | Spain (Boadilla/La Senia) | E3.1 | 2x | Olive genome project |
| var. *europaea* cv. 'Arbequina' | Spain | E1.1 | 2x | Olive genome project |
| var. *europaea* cv. 'Picual' | Spain | E1.1 | 2x | Olive genome project |
| var. *europaea* cv. 'Beladi' | Lebanon | E1.1 | 2x | Olive genome project |
| var. *europaea* cv. 'Sorani' | Syria | E1.1 | 2x | Olive genome project |
| var. *europaea* cv. 'Manzanilla' | Spain | E1.1 | 2x | NCBI (FN996972) |
| var. *europaea* cv. 'Frantoio' | Italy | E1.1 | 2x | NCBI (GU931818) |
| var. *europaea* cv. 'Bianchera' | Italy | E1.1 | 2x | NCBI (NC_013707) |
| var. *sylvestris* | Spain (Pechón) | E3 | 2x | Olive genome project |
| var. *sylvestris* 'Stavrovouni 11' | Cyprus | E1.4 | 2x | NCBI (HF558645) |
| var. *sylvestris* 'Haut Atlas 1' | Morocco (High Atlas) | E2 | 2x | NCBI (NC_015401) |
| var. *sylvestris* 'Gue de Constantine 20' | Algeria:Gue de Constantine, Algiers | E3 | 2x | NCBI (FN997651) |
| *maroccana* | Morocco (Agadir) | M | 6x | Olive genome project |
| *maroccana* 'Immouzzer S1' | Morocco (High Atlas) | M | 6x | NCBI (NC_015623) |
| *cerasiformis* | Portugal (Madeira) | M | 4x | Olive genome project |
| *guanchica* | Spain (Tenerife) | M | 2x | Olive genome project |
| *laperrinei* | Sahara | E1.1 | 2x | Olive genome project |
| *cuspidata*–R | Reunion island | A | 2x | Olive genome project |
| *cuspidata*–S | South Africa | A | 2x | Olive genome project |
| *cuspidata*–I | Iran | C1 | 2x | Olive genome project |
| *cuspidata* 'Almihwit C5.1' | Yemen | C2 | 2x | NCBI (FN996943) |
| *cuspidata* 'Guanghzou 1' | China | C1 | 2x | NCBI (FN996944) |
| *cuspidata* 'Maui 1' | USA (Hawaii-Maui) | A | 2x | NCBI (NC_015604) |

**Table S4.2:** General characteristics of the plastid and mitochondrial genomes of the cv. 'Farga'.

| | Chloroplast | Mitochondria |
|---|---|---|
| Genome size | 136,336 | 593,378 |
| Contigs | 1 | 3 |
| Contigs >1000 | 1 | 2 |
| N50 | 136,336 | 441,284 |
| GC content | 0.37 | 0.45 |
| Number of N | 1,888 | 71,147 |

**Table S4.3:** Admixture coefficient (Q) of each individual per cluster. This table was used to create the Figure 4.4

| | Inferred Clusters | | |
|---|---|---|---|
| *O. europaea* subsp. | 1 | 2 | 3 |
| var. *europaea* 'Farga' | 0.925 | 0.075 | 0.000 |
| var. *europaea* 'Arbequina' | 0.871 | 0.129 | 0.000 |
| var. *europaea* 'Picual' | 0.917 | 0.083 | 0.000 |
| var. *europaea* 'Beladi' | 0.969 | 0.031 | 0.000 |
| var. *europaea* 'Sorani' | 0.959 | 0.041 | 0.000 |
| var. *sylvestris* | 0.641 | 0.359 | 0.000 |
| *maroccana* | 0.420 | 0.580 | 0.000 |
| *cerasiformis* | 0.440 | 0.560 | 0.000 |
| *guanchica* | 0.511 | 0.489 | 0.000 |
| *laperrinei* | 0.412 | 0.588 | 0.000 |
| *cuspidata*–R | 0.000 | 0.000 | 1.000 |
| *cuspidata*–S | 0.000 | 0.039 | 0.961 |
| *cuspidata*–I | 0.253 | 0.411 | 0.335 |

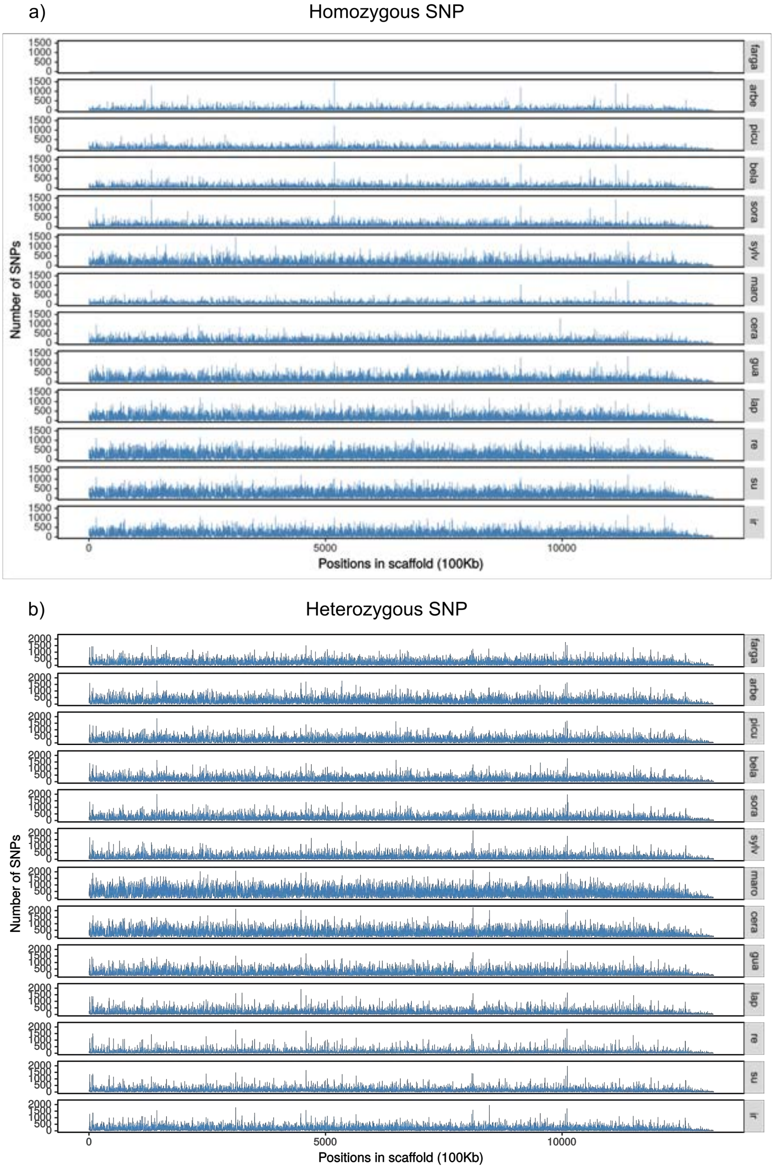

**Figure S4.1:** SNP distribution along the nuclear genome in windows of 100 Kb. a) homozygous SNPs, b) heterozygous SNPs.

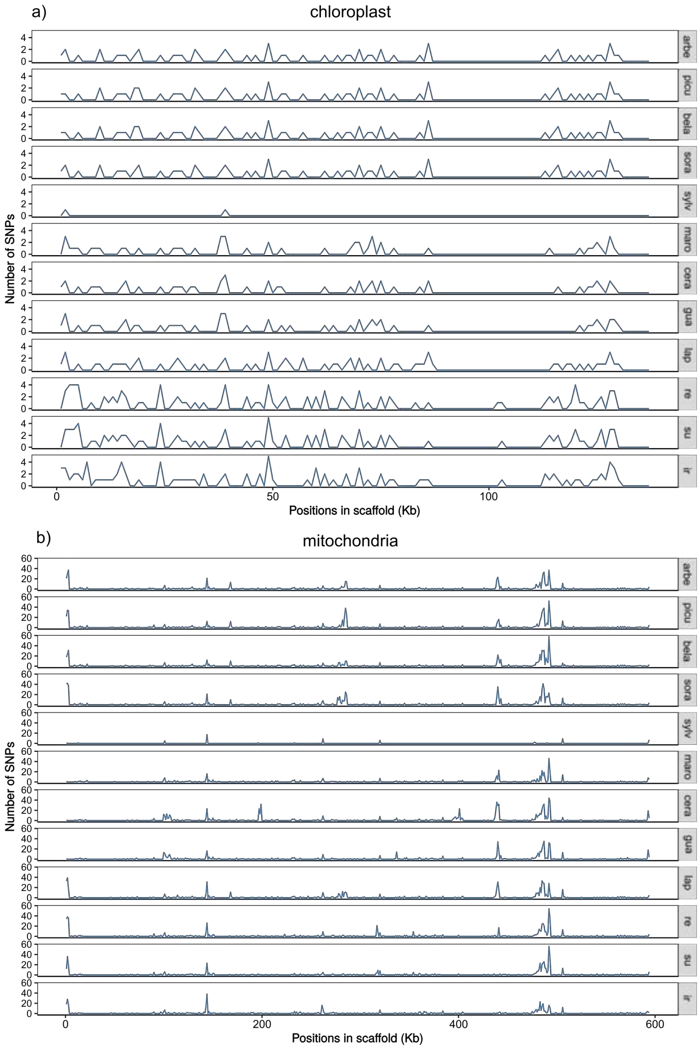

**Figure S4.2:** SNP distribution in the organellar genomes in windows of 1 Kb. a) in the plastid, b) in the mitochondria.

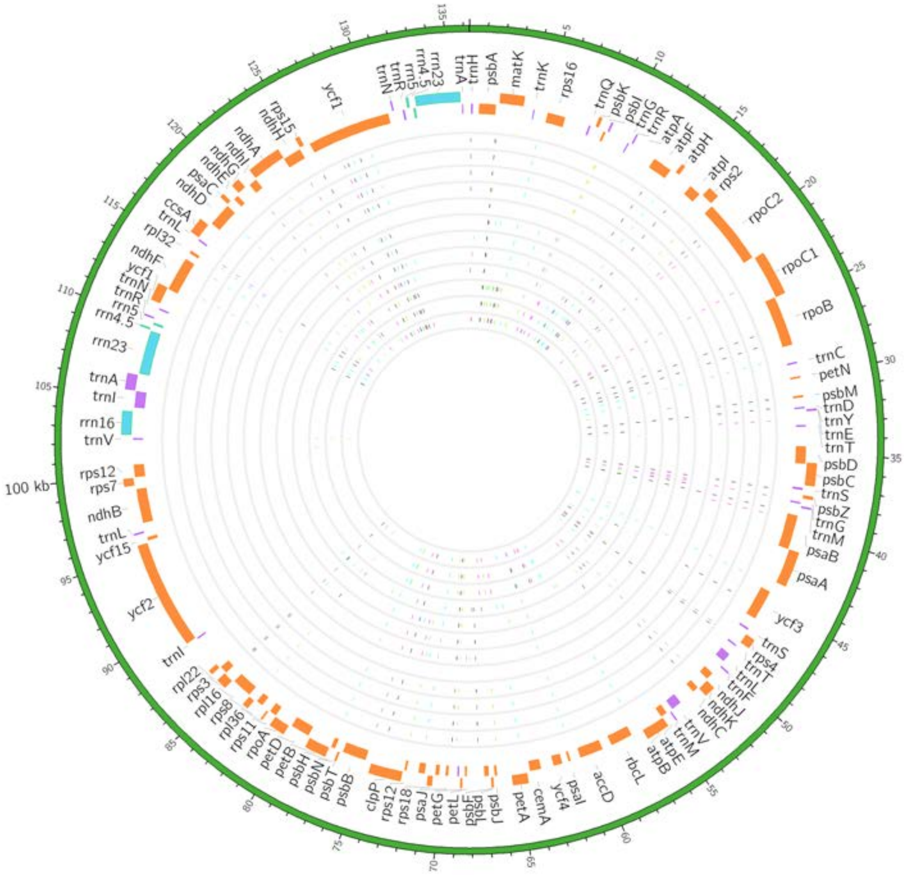

**Figure S4.3:** Partial plastid genome of the cv. 'Farga'. Protein coding genes are shown in orange, rRNAs in blue, and tRNAs in purple. The SNPs are shown per each individual included in this study in the following order starting from outside: 'Arbequina', 'Picual', 'Beladi', 'Sorani', *sylvestris*, *maroccana*, *cerasiformis*, *guanchica*, *laperrinei*, *cuspidata*–R, *cuspidata*–S, *cuspidata*–I.

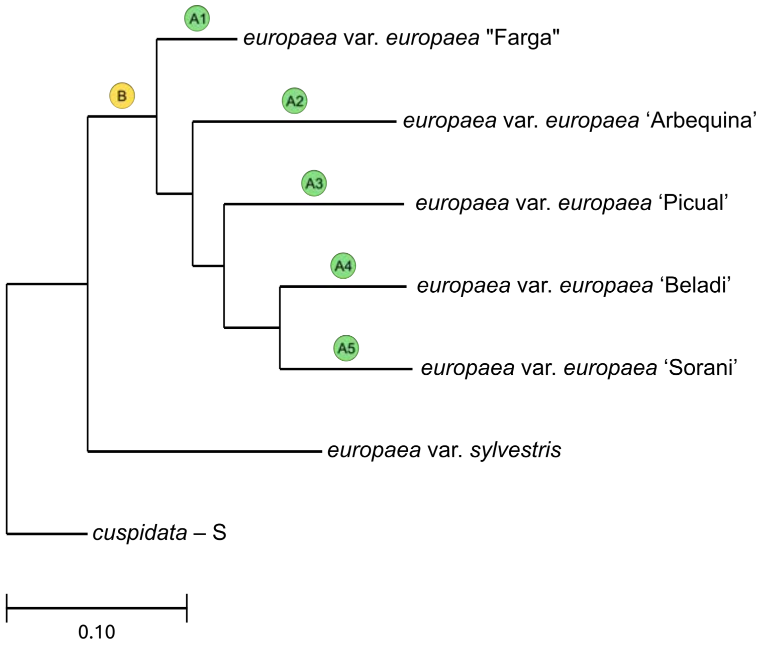
A1
*europaea* var. *europaea* "Farga"
B
A2
*europaea* var. *europaea* ‘Arbequina’
A3
*europaea* var. *europaea* ‘Picual’
A4
*europaea* var. *europaea* ‘Beladi’
A5
*europaea* var. *europaea* ‘Sorani’
*europaea* var. *sylvestris*
*cuspidata* – S
0.10

**Figure S4.4:** Tree used for the tests of selection using the branch-sites model implemented in codeml PAML package. Letters show which branches were marked as having a specific rate, per each of the following hypotheses: A) the terminal branch of each cultivar was marked independently. B) the branch corresponding to the common ancestor of the cultivars.

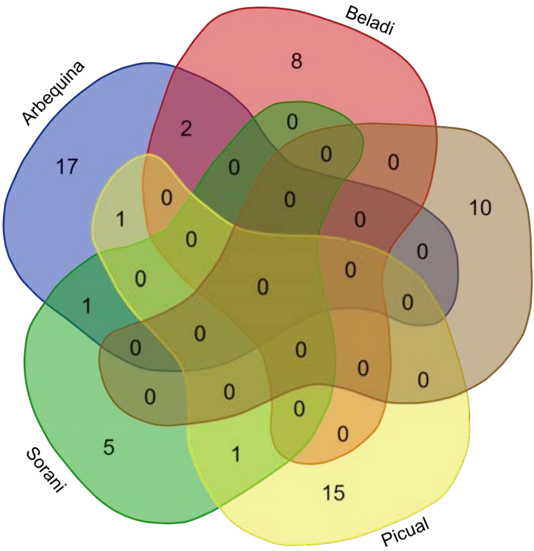

**Figure S4.5:** Venn diagram of the proteins selected per each cultivar in the context of the hypothesis A.

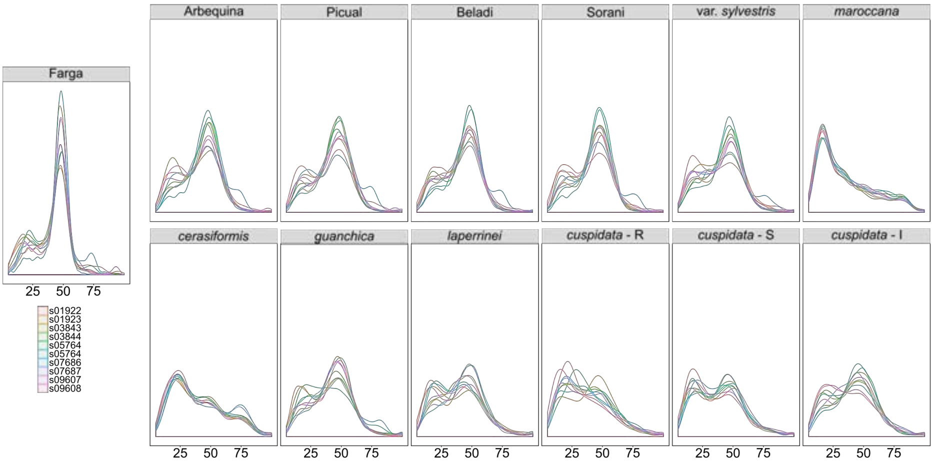

**Figure S4.6:** Density plot of the relative coverage of alternative alleles in heterozygous sites for the 10 larger scaffolds.

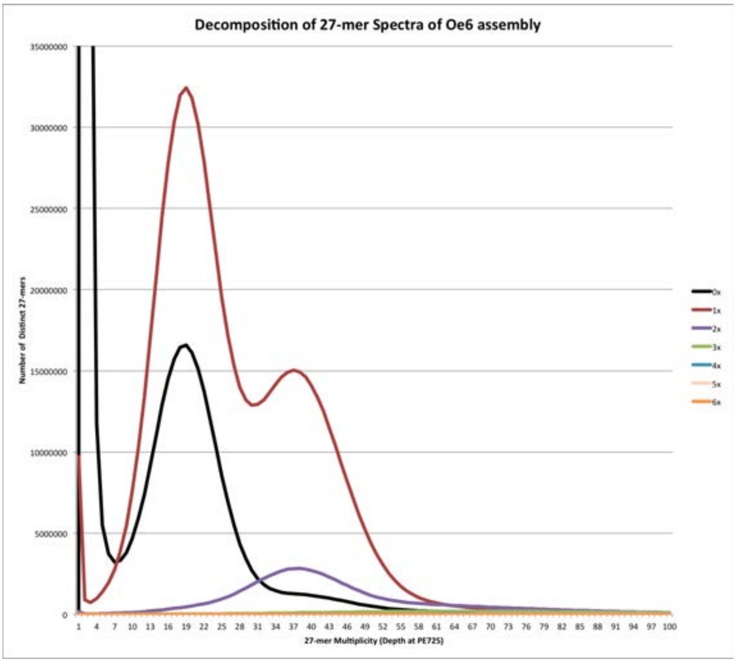

**Figure S4.7:** Plot based on the decomposition of the 27-mer spectra.

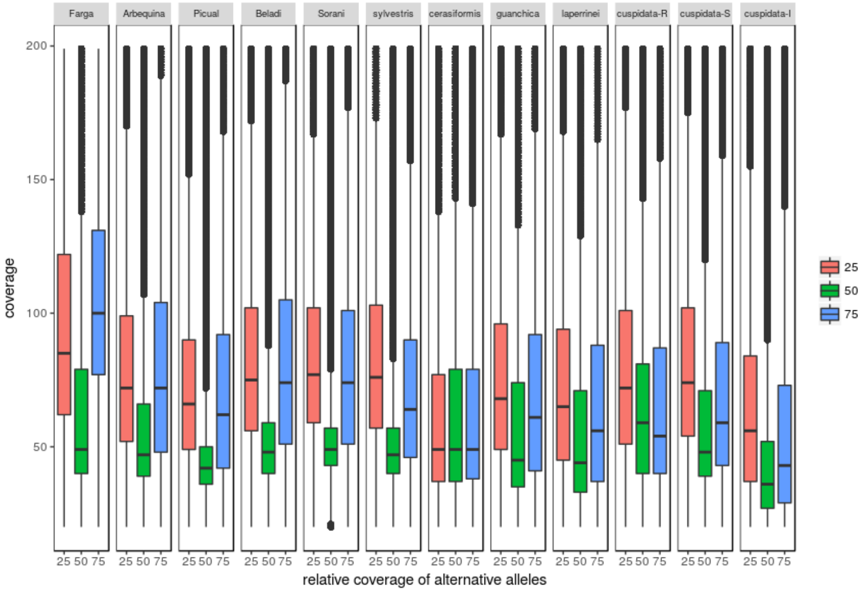

**Figure S4.8:** Box plot of the coverage of the positions with frequencies of 0.25, 0.50, 0.75.

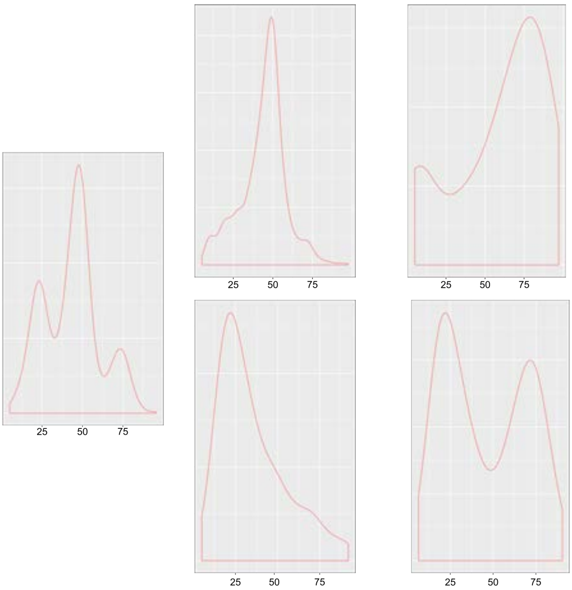

**Figure S4.9:** Density plot of the relative coverage of alternative alleles in heterozygous sites in some fosmid regions.

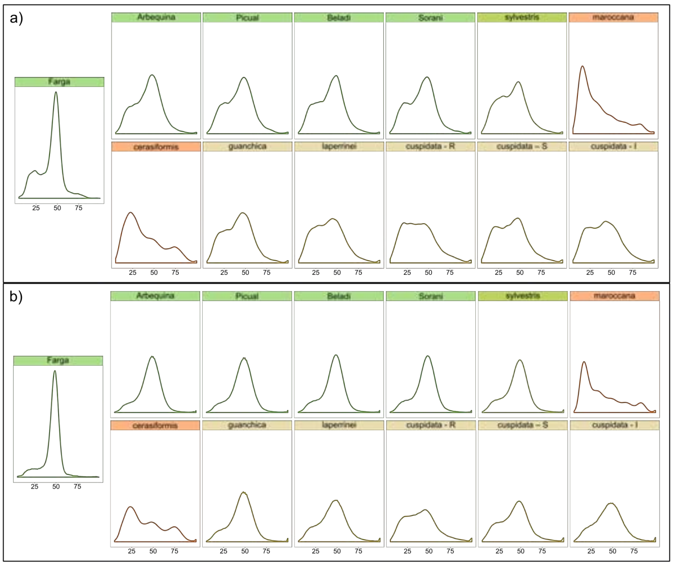

**Figure S4.10:** Density plot for the relative coverage of alternative alleles in heterozygous sites per each individual. a) duplicated genes; b) non-duplicated genes. The cultivated olives are marked on green, *cerasiformis* and *maroccana* in orange, and the other subspecies in yellow.

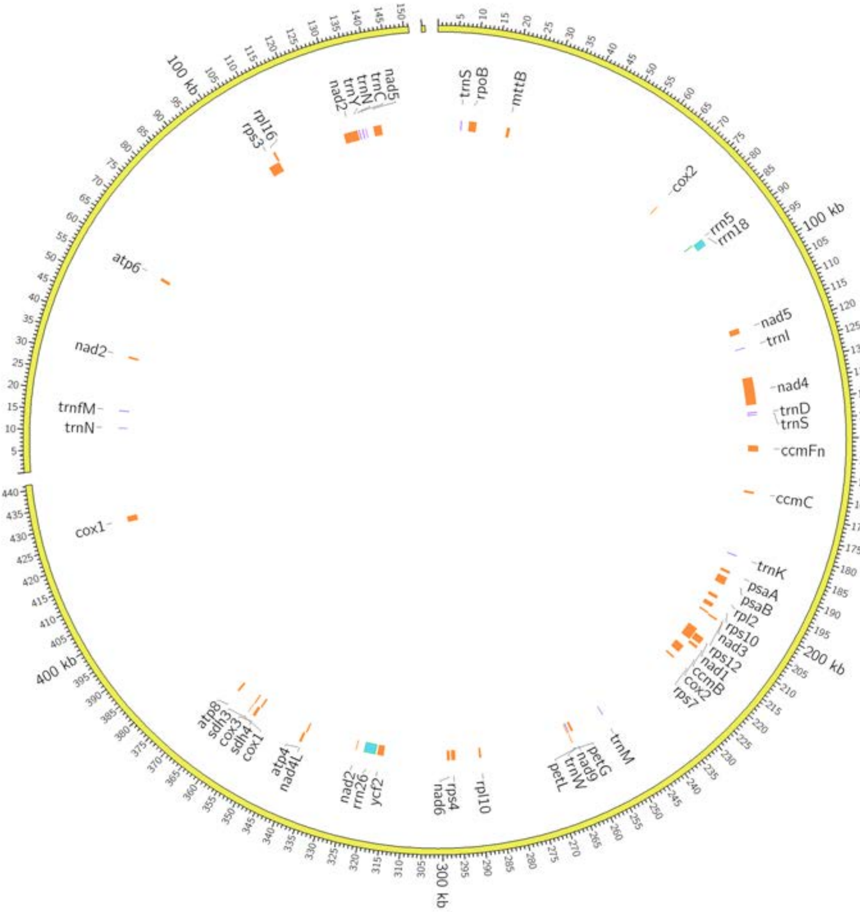

**Figure S4.11:** Partial mitochondrial genome of the cv. 'Farga'. Protein coding genes are shown in orange, rRNAs in blue, and tRNAs in purple.

# 5 Summarizing Discussion

## 5.1 Olive genome

In chapter 2, we presented the assembly and annotation of the first reference genome of olive. It was sequenced from a single individual that is estimated to have been planted around the end of the eighth century (Antonio Prieto-Rodríguez personal communication): *O. europaea* cv. 'Farga'. The final assembly was of 1.32 Gb, which is an intermediate size among other flowering plant genomes. In particular, angiosperm genomes range from 61 Mb of *Genlisia tuberosa* (Fleischmann et al., 2014) to 148.8 Gb of *Paris japonica* (Pellicer et al., 2010). The number of protein coding genes in olive was 56,349. Surprisingly, this number is clearly larger than that of other Lamiales species: *Mimulus guttatus* (28,140 proteins) (Hellsten et al., 2013), *Sesamum indicum* (27,148) (Wang et al., 2014, 2016a), *Utricularia gibba* (26,457) (Ibarra-Laclette et al., 2013), and *Salvia miltiorrhiza* (27,986) (Zhang et al., 2015a). Furthermore, when we compared the proteome of *O. europaea* and that of *M. guttatus* we found that 80.5% of the proteins with at least one homologous gene in *M. guttatus*, had a closer paralogous gene in olive. All these results showed evidence for a WGD in the history of olive, which probably did not involve the other Lamiales species. This result motivated the study presented in Chapter 3.

The availability of the olive genome is an essential resource not only for understanding the evolutionary history of Mediterranean trees, but also for facilitating genetic research and future breeding efforts in this important crop. Climate change, management needs, and the emergence and spread of pathogens are some of the aspects to be palliated generating newer olive cultivars. One way of accelerating this process in this slow-growing tree is the availability of genetic markers associated with traits of agricultural importance (e.g. resistance to pathogens, adaptation to dry environments, etc). In this century, a high number of genetic and molecular tools have been developed in olive, including genetic maps (Besnard et al., 2003; Wu et al., 2004; El Aabidine et al., 2010; Khadari et al., 2010; Marchese et al., 2016; İpek et al., 2017), expressed sequence tags (ESTs) (Alagna et al., 2009; Ozgenturk et al., 2010), and transcriptomes (Galla et al., 2009; Muñoz-Mérida et al., 2013; Bazakos et al., 2015; Guerra et al., 2015; Carmona et al., 2015;

Iaria et al., 2016). In this context the olive genome completes this repertoire of genomic tools and provides a valuable resource for the study of key phenotypic traits. Some earlier studies already detected repeated sequences, mainly five families of tandem repeats (Oe80, Oe86, Oe178, Oe179 and Oe218) and transposable elements (especially long terminal repeat (LTR) retrotransposons) which showed the large repetitive nature of the olive genome (Barghini et al., 2014, 2015). A more recent study identified 227 putative short interspersed nuclear elements (SINEs) and the comparison with other LTR retrotransposon families suggested that the expansion of SINEs in the genome occurred in very ancient times, preceding LTR expansion, and presumably before the separation of Rosids from Asterids (Barghini et al., 2017).

Understanding the genetic basis of the principal biological pathways underlying relevant agricultural traits can be very helpful for the improvement of the productivity, resistance, and nutritional characteristics of crops. However, genetic improvement is difficult in the olive due to its long juvenile phase that ranges from 10 to 15 years (Bracci et al., 2011). Still, many tools have been developed to assess genetic variation in the olive tree. Among these tools we can mention inter-simple sequence repeats (ISSRs) (Terzopoulos et al., 2005; Pasqualone et al., 2001; Vargas and Kadereit, 2001), simple sequence repeats (SSRs) (Rallo et al., 2000; Cipriani et al., 2002; Díaz et al., 2006), random amplified polymorphic DNA (RAPD) (Fabbri et al., 1995; Belaj et al., 2001), amplified fragment length polymorphism (AFLP) (Angiolillo et al., 1999; Pafundo et al., 2005; Rubio de Casas et al., 2006), internal transcribed spacer (ITS) (Hess et al., 2000; Besnard et al., 2007b), plastid sequences (Vargas and Kadereit, 2001) and single nucleotide polymorphisms (SNPs) (Reale et al., 2006; Consolandi et al., 2007; Kaya et al., 2013; Marchese et al., 2016). They have been used alone or in combination for the study of genetic diversity, the characterization of core collections, and for the analysis of the phylogenetic limits in the *O. europaea* complex (Angiolillo et al., 1999; Rubio de Casas et al., 2006; Besnard et al., 2009; Marchese et al., 2016). Nevertheless, more accurate genetic markers are needed for breeding programs.

The availability of the olive genome can facilitate the development of new markers and thus speed up breeding programs, as was previously done with other crops. For example, when the genome of rice became available, multiple authors (Goff, 2002; Yu, 2002) used it to help elucidate the major quantitative trait loci (QTLs) that increases grain productivity. Earlier studies have shown that the gene *OsCKX2* (cytokinin oxidase 2) is the responsible for this QTL (Ashikari, 2005). More recent studies have found that this gene is directly regulated by a zinc finger transcription factor DST (drought and salt tolerance) (Li et al., 2013), which also regulates drought and salt tolerance in rice (Huang et al., 2009). In grape, the genome (Jaillon et al., 2007) facilitated the development of SNPs markers with a high discriminative power for cultivar identification (Cabezas et al., 2011). In tomato, the genome sequence (Tomato Genome Consortium, 2012) helped with the identification of genes that control key agronomic traits. It was possible to identify the esterase (*SlCXE1*) responsible for differences of ester volatile content in tomato fruits (Goulet et al., 2012), a transcription factor (*GLK2*, Golden 2-like) that regulates the photosynthetic capacity of fruits and thus the sugar content (Powell et al., 2012). After the sequencing of the *Cucumis melo* (melon) genome (Garcia-Mas et al., 2012), many QTLs were identified, among which some were associated to phenotypes such as resistance to the cucumber mosaic cucumovirus (CMV) (Guiu-Aragonés et al., 2014), ethylene biosynthesis (Vegas et al., 2013), fruit shape, fruit size, or pulp content (Díaz et al., 2014, 2017). Also the melon genome allowed the identification of the gene that controls iron uptake (*bHLH38*) (Ramamurthy and Waters, 2017) and several candidate genes associated to powdery mildew resistance (Li et al., 2017). In *Prunus persica* (peach), after the release of the genome (International Peach Genome Initiative et al., 2013) it was possible to identify a region on chromosome 1, strongly associated with brown rot resistance and at least two candidate genes associated with pathogen resistance (Martínez-García et al., 2013). Also, it was possible to identify the gene that controls white and yellow coloration of fruit flesh and leaf midvein (carotenoid cleavage dioxygenase, *CCD4*) (Ma et al., 2014), the gene responsible for trichome formation on fruit skin (*PpeMYB25*) which determines the presence or absence of skin pubescence (fuzziness) (Vendramin et al., 2014), and a candidate gene that controls the flat shape of fruits (*PRUPE.6G281100*) (López-Girona et al., 2017). Moreover, the identification of SNP markers tightly associated with six major genes in peach: fruit flesh color *Y*, fruit skin pubescence *G*, fruit shape *S*, sub-acid fruit *D*, stone adhesion-flesh texture *F-M*, and resistance to green peach aphid *Rm2* (Lambert et al., 2016). In *Symphonia globulifera* the availability of the draft genome allowed the development of robust and widely applicable genetic markers (Olsson et al., 2017). More recently, it was shown that the genome sequence of *Mentha longifolia* is a valuable resource for the genetic amelioration of mint cultivars (Vining et al., 2017). This latter study clearly illustrates how the genome sequence can be employed for both metabolic engineering and molecular breeding in pepermint cultivars (Vining et al., 2017). As in other crops, the olive genome together with complementary genomic tools can be used to develop more efficient genetic markers. These markers can be integrated with traditional methods of selection by applying marker-assisted selection (MAS). MAS is generally applied in order to reduce the size of plant populations used during selection through an early selection of genetically predisposed individuals. For example, in olive the use of MAS could save much time and cost allowing a faster identification of beneficial agronomic traits and compatibility relationships in crosses, and a preselection from the progeny of individuals with key agronomic characteristics in an early stage (Díaz, 2012; Sebastiani and Busconi, 2017).

The appearance of new phytopathogens and the restricted availability of suitable environments for the development of the agriculture entails a challenge for breeding programs. Olive unfortunately offers a clear example. In 2013, an outbreak of *Xylella fastidiosa*, causal agent of olive quick decline syndrome (OQDS) once restricted to the Americas, was found in south-eastern Italy (Apulia) (Saponari et al., 2013). Since this initial outbreak, the disease has spread through the majority of the olive trees in the province of Lecce and now represent a threat for the entire European Union (Bosso et al., 2016; Martelli et al., 2016). New genetic markers can be used for mapping genes associated with OQDS-resistance, and thus facilitate marker-assisted selection (MAS) for resistant olives.

Target genetic markers can also be used for identification purposes. In the olive, SSRs are widely used to distinguish among different cultivars and have resolved some cases of homonyms and synonyms (Lopes et al., 2004; Omrani-Sabbaghi et al., 2007; Bracci et al., 2009; Abdessemed et al., 2015).

Indeed, the availability of more markers may lead to the finding of more of these cases among olive cultivars. Also, these markers can be useful for the traceability of the olive oil, avoiding deliberate or accidental mislabelling (Bracci et al., 2011).

Remarkably, the availability of the olive genome opens a wide spectrum of opportunities to understand its biology and genetic diversity. Moreover, the genome sequence of this monumental ancient tree (cv. 'Farga') is a good genetic source since it has survived local biotic and abiotic stresses for centuries.

## 5.2 Polyploidy in *Olea europaea*

The results found in Chapter 2 suggested the presence of at least one WGD in the history of *O. europaea*. In chapter 3 we used phylogenomic analysis to investigate past genome duplications in olive, and found that the olive underwent three polyploidization events since the divergence from the other non-Oleaceae Lamiales lineages included in this study. The first one likely occurred at the base of the family Oleaceae and was described as an allopolyploidization between a species more similar to the other non-Oleaceae Lamiales, and another one that diverged earlier. We also demonstrated that this WGD is different from the one previously described at the base of the non-Oleacea Lamiales (Ibarra-Laclette et al., 2013), and that it dates back to 33–72 millions years ago (Mya). Interestingly, this date was correlated with the Cretaceous–Paleogene (K/Pg) boundary (66 Mya), the most recent mass extinction event, where many other plant species have undergone a WGD (Lohaus and Van de Peer, 2016; Vanneste et al., 2014; Van de Peer et al., 2017). The second wave of duplications was placed at the basal node of the tribe Oleeae and was described also as an allopolyploidization event. In this case our results suggested that one of the parental plants was a close relative of jasmine species. Recent studies have shown that *O. europaea* and *Fraxinus excelsior* shared a WGD placed at a deep node of the family Oleaceae (Sollars et al., 2016), but our analysis suggested that this event involves the ancestor of the tribe Oleeae. Our results of further duplications are also supported by the fact that the chromosome numbers in the other tribes of Oleaceae (Myxopyreae, Forsythieae, Fontanesieae, Jasmineae) are generally lower (2n = 22, 28 with some exceptions) than those in the Oleeae tribe (2n = 46) (Taylor, 1945). Interestingly, earlier studies suggested that the tribe Oleeae (whole 23-chromosome group) had an allopolyploid origin as the result of a hybridization between two unknown and now probably extinct species of the other tribes with x = 11 and 12 (Taylor, 1945). More recent studies showed that the chromosome morphology and the pattern of heterochromatic bands in *O. europaea* have signs of interspecific hybridization followed by chromosome doubling (Falistocco and Tosti, 1996). In chapter 3 we propose that the chromosome number 2n = 46 in the tribe Oleeae, came from a base number of 12 (x = 12). This means that it was some type of chromosome fusion or loss during the polyploidization process resulting in an ancestor with 23 chromosomes. Particularly in *O. europaea*, the chromosome pair XIV bears heterochromatin of 'telomeric' type around the centromere, which was proposed as evidence of end-to-end fusion of two chromosomes (Minelli et al., 2000). Based on this result they inferred that the proposed diploid chromosome number of 2n = 46 found in several genera of Oleaceae is the result of of an ancestor of 2n = 48 (Minelli et al., 2000). This chromosome number is observed in certain species of the genus *Syringa*, which belongs to the same tribe of the olive (Taylor, 1945). Assuming that the diploid ancestor of the tribe Oleeae had chromosome rearrangements resulting in a change from 2n = 24 to 2n = 23 through loss or fusion of chromosomes, we can accept the possibility of an independent WGD in the tribe Oleeae. The date of this WGD is placed around 14–33 Mya, which corresponds with the Oligocene and Miocene periods. This period of time is characterized by two glaciations (Oi-1 - 34 Mya, and Mi-1 - 23 Mya) and the middle Miocene climate optimum (17 to 15 Mya), one of the warmest phases since the Miocene (Zachos, 2001; Zachos et al., 2001), and is accompanied by accelerated rates of diversification (Vargas et al., 2014; Hinsinger et al., 2013; Divakar et al., 2012; Kürschner and Kvaček, 2009). The third event is specific to *O. europaea*; however, the data at hand cannot be used to distinguish between auto- or allopolyploidization processes at this level. Additional genomes sequenced from closer relatives (i.e. other species of *Olea* and closely related genera) would be needed to assess such difference.

Remarkably, our analysis highlighted the versatility of phylomes as a phylogenomic tool for the detection of polyploidization and for the differentiation between allo- and autopolyploidization. Many studies in plants have used synteny and phylogenomics for the estimation of polyploidy and the principal phylogenomic method used is age distribution. These tools were applied to detect ancient polyploidy in angiosperms, in seed plants (Jiao et al., 2011), in monocots (Jiao et al., 2014), in core eudicots (Jiao et al., 2012), in Poales (McKain et al., 2016), and more recent WGDs in *Vigna* (Kang et al., 2014), *Brassica* (Liu et al., 2014a; Wang et al., 2011), *Musa acuminata* (D'Hont et al., 2012), *Gossypium arboreum* (Li et al., 2014), *Utricularia gibba* (Ibarra-Laclette et al., 2013), among others. However, the analysis of phylomes alone or in combination with other tools offers a different approach for the estimation of polyploidy and the possibility of telling apart allo- and autopolyploidization. Phylome analysis of *Phaseolus vulgaris* allowed the identification of a WGD at the base of the subfamily Papilionoideae (Vlasova et al., 2016). In *Saccharomyces cerevisiae* similar analyses revealed its allopolyploid origin (Marcet-Houben and Gabaldón, 2015). Furthermore, chapter 3 includes, for the first time, a different approach of the 4DTv analysis. This method is generally used by plotting all the set of paralogs proteins of any species, eg. *Prunus persica* (International Peach Genome Initiative et al., 2013), *Salix suchowensis* (Dai et al., 2014), *Punica granatum* (Qin et al., 2017), *Cucurbita pepo* (Montero-Pau et al., 2017), *Lepidium meyenii* (Zhang et al., 2016). In chapter 3 we show how the analysis of the 4DTv can be done using sets of paralogs that occurred at different evolutionary ages as extracted from the phylome and how it can be used to distinguish auto- and allopolyplidization events.

An additional polyploidization (or partial genome duplication) is proposed in chapter 4 as a result of our analysis of relative coverage of alternative alleles in heterozygous sites of the *Olea europaea* complex. This kind of analysis is generally used to detect polyploidy in specific cells or tissues, including detection of cancer cases in humans (Lundberg et al., 2013; Rodríguez-Santiago et al., 2010; Cutcutache et al., 2016; Van Loo et al., 2010) or estimation of ploidy levels in *Malus domestica* (Chagné et al., 2015), *Saccharomyces cerevisiae* (Corrêa dos Santos et al., 2017; Weiß et al., 2017), and

*Phytophthora infestans* (Weiß et al., 2017).

This polyploidization event should have taken place after the last WGD described in chapter 3 and involving the ancestor that diverged into the subspecies of the *O. europaea* complex. It has been consistently inferred using different molecular tools that subsp. *cuspidata* is the first group that diverged (8.3–4.0 Mya) within the complex (Vargas and Kadereit, 2001; Rubio de Casas et al., 2006; Besnard et al., 2009). In particular, this last WGD should be placed earlier than this period and thus earlier than the domestication event (6,000 years ago). In other words, the most recent common ancestor of the *O. europaea* complex may have been already a neopolyploid plant. In general, our results from chapter 3 and 4 show signs of two independent WGDs in the ancestor of *O. europaea*, which should be placed after its divergence from *Phillyrea angustifolia* and before the divergence of the different subspecies of the *O. europaea* complex. However, the fact that the closer species included in this analysis (*Phillyrea angustifolia*) does not share these events but presents the same number of chromosomes (2n = 46), as indicated by cytological studies, is at odds.

*O. europaea* with 1.32 Gb of genome size and 2n = 46 shows, according to our results, a complex scenario of up to four polyploidization events since an ancestor of Lamiales (33–72 Mya). This number of polyploidization events is not surprising if we take into consideration that other plant species with smaller genome sizes and chromosome numbers have many WGDs in their evolutionary history. For instance, *Utricularia gibba*, with a genome of 82 Mb and 2n = 28, underwent three rounds of WGDs since the divergence form the family Oleaceae (62-72 Mya) (Ibarra-Laclette et al., 2013). *Musa acuminata*, with 520 Mb and 2n = 22 has two WGDs since its divergence from *Zingiber officinale* (∼70 Mya) (D'Hont et al., 2012). Interestingly these two species have the same chromosome number 2n = 22 (Das et al., 1998), despite the specific WGDs in only one of the lineages. *Arabidopsis thaliana* with 117 Mb and 2n = 10 underwent two WGDs after its divergence from *Carica papaya* (∼70 Mya) (Van de Peer et al., 2009a), which has a higher number of chromosomes (2n =18) and after this period did not underwent any WGD (Ming et al., 2008). Thus, we consider that the fact that *O. europaea* has the same number than *P. angustifolia* does not invalidate our results. Several scenarios could reconcile both results including independent, parallel events in the *P. angustifolia* lineage or fast diploidization and genome rearrangements following WGD in *O. europaea*. Nevertheless further research, including other closely related species, is needed.

In summary, our results show signatures of at least four genome duplication events within Lamiales in the evolutionary history of the ancestor of the *O. europaea* complex. In addition, two more events of polyploidization increase within the *Olea europaea* complex have been documented in the past based on chromosome numbers: tetraploid (2n = 92, subsp. *cerasiformis*) and hexaploid (2n = 138, subsp. *maroccana*) (Besnard et al., 2008). Some of these events appear to be associated with climatic changes and mass extinction events (Van de Peer et al., 2017); however, a more careful analysis of the dates would be necessary to further explore these associations. All these results lead us to consider a WGD for the ancestor of the *Olea europaea* before domestication, which supports the idea that domestication followed polyploidization (Salman-Minkov et al., 2016; Fang and Morrell, 2016). However, the role of polyploidization in the domestication of the olive needs to consider a more complex scenario because the most important domestication process in the *Olea europaea* complex involves the subspecies *europaea* (2n = 46), but did not involve the tetraploid and hexaploid subspecies.

## 5.3 Hybridization of the *O. europaea* complex

In chapter 4 we presented the first phylogenomic analysis of the *O. europaea* complex and we showed that the evolution of the different lineages was shaped by extensive genetic flux as a product of frequent hybridization. These results corroborate the observation made in previous studies based on nuclear and organellar markers (Rubio de Casas et al., 2006; Besnard et al., 2009, 2007b; Besnard and El Bakkali, 2014). Moreover, *O. europaea* has an allogamous mode of reproduction (Besnard et al., 2000; Breton et al., 2017; Mookerjee et al., 2005) and the subspecies are interfertile, which promotes hybridization (Besnard et al., 2001b; Contento et al., 2002; Besnard et al., 2009; Cáceres et al., 2015; Mousavi et al., 2017; Besnard and El Bakkali, 2014).

An even more important role of hybridization is herein proposed for evolution and divergence of *O. europaea*. Because different ploidy levels have been observed in the *O. europaea* complex (Besnard et al., 2008; Besnard and Baali-Cherif, 2009), two kinds of hybridization could have taken place, namely homoploid (with no variation of the ploidy level) and allopolyploid hypbridization (varying the ploidy level). Furthermore, hybridization is a common process in plants and a source for genetic variance in a few generations (Rieseberg, 1997; Renaut et al., 2014). In potatoes homoploid hybridization has been proposed as the main mechanism involved in the origin and evolution of the diploid species (Masuelli et al., 2009). Furthermore, intraspecific hybridization can be a powerful process impacting the evolution of invasiveness in certain species, such as *Pyrus calleryana* (Culley and Hardiman, 2009) and *Schinus terebinthifolius* (Williams et al., 2007). In this context two subspecies of *O. europaea* (*europaea* and *cuspidata*) have successfully invaded several regions in Australia, New Zealand and Pacific islands, and hybridization between these two lineages has been proposed as an important process during the olive invasion (Besnard et al., 2007a, 2014). Hybridization, also can provide the necessary genetic variation for adaptive evolution within a species, as was previously shown in the recovery of fitness in *Chamaecrista fasciculata* (Erickson et al., 2006) and the improvement of the production in *Jatropha curcas* (Ayizannon et al., 2017).

Genetic improvement of the olive through artificial intraspecific hybridization has been recently discussed (Cáceres et al., 2015; Rugini et al., 2016). An increasing demand of olive products leads to promote a higher density of orchards, and for that we need trees (genotypes) with reduced size, reduced apical dominance, a semi-erect growth habit, easy to propagate, resistant to abiotic and biotic stresses, with reliably high productivity and quality of both fruits and oil (Rugini et al., 2016). In general, the wild *O. europaea* lineages show more genetic diversity than cultivated olives, as we have shown in chapter 4, which confirms previous results (Lumaret et al., 2004; Belaj et al., 2010; Besnard et al., 2011). Furthermore these wild lineages are adapted to many environmental conditions, including resistance to both abiotic stress, such as drought, salt, wind and low temperature, and biotic stress, such as damages and infections caused by *Verticillium dahliae*, *Spilocaea oleaginea*, *Xylella fastidiosa*, and olive fly (Aranda et al., 2011; Trapero et al., 2015; Ciccarese et al., 2002; Mkize et al., 2008; Hannachi et al., 2009; Rugini et al., 2016; Giampetruzzi et al., 2017). In particular, oleaster (var. *sylvestris*) and subsp. *cuspidata* have been proposed as suitable genetic resources (Hannachi et al., 2009; Sheidai et al., 2010; Beghé et al., 2017). Natural hybrids have been observed between oleaster - cultivars (Zohary and Spiegel-Roy, 1975; Besnard and Bervillé, 2000), and *europaea* - *cuspidata* (Hannachi et al., 2009; Besnard et al., 2007a). Experimental crossing has been performed in order to enrich the germplasm of cultivars. Progeny of crosses between cultivars and oleaster have shown the highest mean values for vigor traits (i.e., tree height and trunk diameter) and short juvenile period when compared with progenies of crosses between cultivars (Klepo et al., 2014). Also crosses between subspp. *europaea* and *cuspidata* have been reported (Besnard et al., 2001b; Cáceres et al., 2015). For this reason an increasing number of studies have been focused on understanding the biology of the subsp. *cuspidata*. For instance the floral biology analysis has shown that the *cuspidata* flower is very similar to that of *europaea* (Caceres et al., 2016), which additionally share the same self-incompatibility system (Breton et al., 2017). However, according to our knowledge similar studies involving other subspecies have not been reported. The implementation of olive breeding programs with the inclusion of the other subspecies (*guanchica*, *laperrinei*, *cerasiformis*, and *maroccana*) might represent a useful strategy to exploit the enormous genetic pool that these lineages harbour. For instance, subsp. *laperrinei* is adapted to the dry conditions that the Sahara dessert provides, subsp. *maroccana* shows evidence of certain degree of domestication and subspp. *guanchica* and *cerasiformis* are particularly well adapted to the mild conditions of oceanic islands.

In summary our results have shown a complex evolutionary history in the *O. europaea* complex where hybridization has been predominant. Our limited tools help identify past genetic interchange among the different lineages, which lead us to infer that a higher number of hybridization events may have taken place in the evolutionary history of the olive tree. Different molecular tools contrasting phylogenetic relationships have already shown that incongruence between the organellar and nuclear trees may be related to processes of hybridization. That is why wider sampling of populations and cultivars could shed further light on the role of hybridization in the domestication of numerous cultivars.

## 5.4 Origin of cv. 'Farga'

Our results presented in chapter 4 have also shown that the cv. 'Farga' (our reference genome) has a different phylogenetic history with respect to the other cultivars included in this study. It is highlighted by the incongruence observed between the three-genome phylogenetic trees. In the nuclear tree 'Farga' is close to the other cultivars, while in the plastid and mitochondrial trees 'Farga' is closer to the var. *sylvestris* from Pechón (Spain), that was sequenced in this project. This incongruence suggests that a hybridization event between an individual closely related to var. *sylvestris*, as ovule donor, and a domesticated individual, acting as pollen donor, gave rise to the cv. 'Farga'. It is also noteworthy that cv. 'Farga' and the var. *sylvestris* have a rare chlorotype E3 with origin in the western area (Besnard et al., 2013b). These results support the hypothesis of a secondary domestication event in the western Mediterranean basin, where wild native olives interbred with cultivated olives that probably came from the first domestication event in the east of the Mediterranean basin (Besnard et al., 2013a,b).

## 5.5 Gene selection in olive

In chapter 4 we have also shown that different sets of proteins are likely under positive selection in each cultivar. Some of these proteins are associated with response to biotic and abiotic stress, and with metabolic and developmental processes, while others have unknown function. Plants are constantly exposed to a broad range of environmental stresses. Particularly, crops in the field are affected by drought, salinity, heat, cold, chilling, freezing, nutrient, high light intensity, ozone (O3), anaerobic stress, phytopathogens and pests (Suzuki et al., 2014). Therefore, it is not surprising that genes under positive selection are associated with environmental response and biotic and abiotic stress tolerance. In addition domestication is a process that involves artificial selection of beneficial traits leading to dramatic alteration of the morphology, physiology, and life history of cultivated plants when compared with the wild progenitors (Darwin, 1859). For instance, in other cultivated plants such as tomato (Koenig et al., 2013), rice (Sun et al., 2015) and ramie (Liu et al., 2014b) positively selected genes associated with domestication process have also been observed.

Our results indicate that domestication may have played a role in selecting key characteristics in the different olive cultivars.

## 5.6 Future perspectives

Genome sequence of woody plants is challenging due to the presence of a large proportion of repeat elements, WGDs, and a high level of nucleotide diversity. Despite all these complications in this work we present the first reference genome for olive. The availability of the olive genome is an important step for the elucidation of its evolutionary history and for the study of genes and molecular mechanisms underlying important agronomic traits. The olive genome presented in this study is the first step that will lead to future genetic amelioration. The increasing development of tools and technologies for genome sequencing and genome assembly will allow the availability of a high-quality phased genome at chromosome level in a near future. This type of genome will clarify many of the scenarios presented in this work such as the polyploidy origin in *O. europaea* and the genetic structure of the cultivars. Moreover, the availability of genomes or transcriptomes of genera closely related to the olive will shed light on the last WGD detected by phylogenomic analysis. The genome sequencing of a higher number of individuals of the different subspecies of the *O. europaea* complex will help to clarify the evolutionary history of subsp. *europaea* and the taxonomic limits of the subsp. *cuspidata*.

The identification of new sequence polymorphisms, e.g. SNPs, can be useful in the development of modern molecular markers, high-throughput genome-wide genotyping, and the implementation of more efficient breeding protocols. In this work we presented one of the largest sets of predicted SNPs for olive, which can further be experimentally validated for this kind of purposes.

In sum, sequencing of a high-quality olive genome from all cultivars and lineages of the *O. europaea* complex will facilitate the use of molecular and genomic tools more extensively in order to better understand the evolution of the wild olive tree, to reconstruct the processes of domestication and to speed up breeding programs.

# 6 Conclusions

The following conclusions provide the main contributions of this PhD thesis.

1. The first whole reference genome sequence of the cultivated olive tree has been assembled and annotated, providing a hallmark resource for the study of this important crop. Alongside the reference genome, twelve other individuals from cultivated and wild subspecies of *O. europaea* have been analysed.

2. Phylogenomics was succesfully used to disentangle past allo- and autopolyploidizations. This approach was used here to differentiate three polyploidization events in the evolutionary history of *O. europaea* since its divergence from the other non-Oleaceae Lamiales included in this study. Two are allopolyploidization events at the base of the family Oleaceae and tribe Oleeae, respectively, and the last one is a whole genome duplication specific to the lineage of Olea where the olive is placed.

3. The patterns of allelic variation suggest an additional polyploidization event in the history of *O. europaea*. This event predates the divergence of the different subspecies of the *O. europaea* complex.

4. Whole genome phylogenies of wild and cultivated olives support recurrent homoploid hybridization.

5. Cultivar ‘Farga’ has a different evolutionary history than the other studied cultivars suggesting crossing of wild plants of the western Mediterranean with a cultivated stock from the eastern Mediterranean.

6. Olive domestication appears to have involved positive selection of genes associated with the response to biotic and abiotic stress and with developmental processes.

In sum, whole genome sequencing helps to reconstruct polyploidization events in the course of evolution of the olive tree as well as to propose rampant hybridization entailing gene duplication.

# 7 Appendix: List of publications

1. Fernando Cruz*, **Irene Julca***, Jèssica Gómez-Garrido, Damian Loska, Marina Marcet-Houben, Emilio Cano, Beatriz Galán, Leonor Frias, Paolo Ribeca, Sophia Derdak, Marta Gut, Manuel Sánchez-Fernández, Jose Luis García, Ivo G. Gut, Pablo Vargas, Tyler S. Alioto and Toni Gabaldón. (2016). Genome sequence of the olive tree, *Olea europaea*. *Gigascience*, 5:29. (*Contributed equally)

2. **Irene Julca***, Marina Marcet-Houben*, Pablo Vargas, and Toni Gabaldón. (2017). Phylogenomics of the olive tree (*Olea europaea*) disentangles ancient allo- and autopolyploidizations in Lamiales. *BMC Biology (submitted)*. (*Contributed equally).

3. **Irene Julca**, Marina Marcet-Houben, Fernando Cruz, Ivo G. Gut, Tyler S. Alioto, Pablo Vargas, and Toni Gabaldón. Genome sequencing of wild and cultivated olive trees reveals rampant hybridization in the *Olea europaea* complex. *In preparation*.

4. **Irene Julca**, Samir Droby, Noa Sela, Marina Marcet-Houben, and Toni Gabaldón. (2016). Contrasting Genomic Diversity in Two Closely Related Postharvest Pathogens: *Penicillium digitatum* and *Penicillium expansum*. *Genome Biology and Evolution*, 8(1):218–227.

5. Thomas C. Mathers, Yazhou Chen, Gemy Kaithakottil, Fabrice Legeai, Sam T. Mugford, Patrice Baa-Puyoulet, Anthony Bretaudeau, Bernardo Clavijo, Stefano Colella, Olivier Collin, Tamas Dalmay, Thomas Derrien, Honglin Feng, Toni Gabaldón, Anna Jordan, **Irene Julca**, Graeme J. Kettles, Krissana Kowitwanich, Dominique Lavenier, Paolo Lenzi, Sara Lopez-Gomollon, Damian Loska, Daniel Mapleson, Florian Maumus, Simon Moxon, Daniel R. G. Price, Akiko Sugio, Manuella van Munster, Marilyne Uzest, Darren Waite, Georg Jander, Denis Tagu, Alex C. C. Wilson, Cock van Oosterhout, David SwarbreckEmail author and Saskia A. Hogenhout. (2017). Rapid transcriptional plasticity of duplicated gene clusters enables a clonally reproducing aphid to colonise diverse plant species. *Genome Biology*, 18:27.

# References

Abdelfattah, A., Li Destri Nicosia, M. G., Cacciola, S. O., Droby, S., and Schena, L. (2015). Metabarcoding Analysis of Fungal Diversity in the Phyllosphere and Carposphere of Olive (Olea europaea). *PloS one*, 10(7):e0131069.

Abdessemed, S., Muzzalupo, I., and Benbouza, H. (2015). Assessment of genetic diversity among Algerian olive (Olea europaea L.) cultivars using SSR marker. *Scientia Horticulturae*, 192:10–20.

Abro, S., Kandhro, M. M., Laghari, S., Arain, M. A., and Deho, Z. A. (2009). Combining ability and heterosis for yield contributing traits in upland cotton (Gossypium Hirsutum L.). *Pakistan Journal of Botany*, 41(4):1769–1774.

Abrouk, M., Murat, F., Pont, C., Messing, J., Jackson, S., Faraut, T., et al. (2010). Palaeogenomics of plants: Synteny-based modelling of extinct ancestors.

Adams, K. L. and Wendel, J. F. (2005). Novel patterns of gene expression in polyploid plants.

Akerborg, O., Sennblad, B., Arvestad, L., and Lagergren, J. (2009). Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 106(14):5714–5719.

Al-Shahrour, F., Díaz-Uriarte, R., and Dopazo, J. (2004). FatiGO: A web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, 20(4):578–580.

Alagna, F., D'Agostino, N., Torchia, L., Servili, M., Rao, R., Pietrella, M., et al. (2009). Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. *BMC Genomics*, 10(1):399.

Alexander, D. H. and Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, 12(1):246.

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664.

Amane, M., Lumaret, R., Hany, V., Ouazzani, N., Debain, C., Vivier, G., and Deguilloux, M. F. (1999). Chloroplast-DNA variation in cultivated and wild olive (Olea europaea L.). *Theoretical and Applied Genetics*, 99(1-2):133–139.

Anderson, E. (1949). Introgressive Hybridization. *Wiley*.

Anderson, E. C. (2008). Bayesian inference of species hybrids using multilocus dominant genetic markers. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 363(1505):2841–50.

Anderson, E. C. and Thompson, E. A. (2002). A model-based method for identifying species hybrids using multilocus genetic data. *Genetics*, 160(3):1217–29.

Angiolillo, A., Mencuccini, M., and Baldoni, L. (1999). Olive genetic diversity assessed using amplified fragment length polymorphisms. *TAG Theoretical and Applied Genetics*, 98(3-4):411–421.

Anis, G. B., Ibrahem El-Sherif, A., Freeg, H., and Arafat, E. F. (2017). Evaluation of some hybrid combinations for exploitation of two line system heterotic in Rice (Oryza sativa L.). *International Journal of BioSciences, Agriculture and Technology*, 8(1s):975–4539.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408(6814):796–815.

Aranda, S., Montes-Borrego, M., Jiménez-Díaz, R. M., and Landa, B. B. (2011). Microbial communities associated with the root system of wild olives (Olea europaea L. subsp. europaea var. sylvestris) are good reservoirs of bacteria with antagonistic potential against Verticillium dahliae. *Plant and Soil*, 343(1-2):329–345.

Arrigo, N. and Barker, M. S. (2012). Rarely successful polyploids and their legacy in plant genomes.

Ashikari, M. (2005). Cytokinin Oxidase Regulates Rice Grain Production. *Science*, 309(5735):741–745.

Aversano, R., Contaldi, F., Ercolano, M. R., Grosso, V., Iorizzo, M., Tatino, F., et al. (2015). The Solanum commersonii Genome Sequence Provides Insights into Adaptation to Stress Conditions and Genome Evolution of Wild Potato Relatives. *The Plant cell*, 27(4):954–68.

Ayizannon, R., Ahoton, L., Ezin, V., Quenum, F., and Mergeai, G. (2017). Improvement of physic nut (Jatropha curcas L.) by intraspecific hybridization between ecotypes of Africa and Americana. *Journal of Plant Breeding and Crop Science*, 9(5):54–62.

Baali-Cherif, D. and Besnard, G. (2005). High genetic diversity and clonal growth in relict populations of Olea europaea subsp. laperrinei (Oleaceae) from Hoggar, Algeria. *Annals of Botany*, 96(5):823–830.

Badouin, H., Gouzy, J., Grassa, C. J., Murat, F., Staton, S. E., Cottret, L., et al. (2017). The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature*, 546(7656):148–152.

Badr, A., M, K., Sch, R., Rabey, H. E., Effgen, S., Ibrahim, H. H., Pozzi, C., Rohde, W., and Salamini, F. (2000). On the Origin and Domestication History of Barley (Hordeum vulgare). *Molecular Biology and Evolution*, 17(4):499–510.

Ballvora, A., Ercolano, M. R., Weiß, J., Meksem, K., Bormann, C. A., Oberhagemann, P., et al. (2002). The R1 gene for potato resistance to late blight (Phytophthora infestans) belongs to the leucine zipper/NBS/LRR class of plant resistance genes. *Plant Journal*, 30(3):361–371.

Bandelj, D., Jakše, J., and Javornik, B. (2004). Assessment of genetic variability of olive varieties by microsatellite and AFLP markers. *Euphytica*, 136(1):93–102.

Bandelt, H. J. and Dress, A. W. (1992). Split decomposition: A new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution*, 1(3):242–252.

Bandelt, H. J., Forster, P., and Rohl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, 16(1):37–48.

Bandelt, H. J., Forster, P., Sykes, B. C., and Richards, M. B. (1995). Mitochondrial portraits of human populations using median networks. *Genetics*, 141(2):743–753.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology*, 19(5):455–77.

Barber, J. C., Finch, C. C., Francisco-Ortega, J., Santos-Guerra, A., and Jansen, R. K. (2007). Hybridization in Macaronesian Sideritis (Lamiaceae): Evidence from incongruence of multiple independent nuclear and chloroplast sequence datasets. *Taxon*, 56(1):74–88.

Barghini, E., Mascagni, F., Natali, L., Giordani, T., and Cavallini, A. (2017). Identification and characterisation of Short Interspersed Nuclear Elements in the olive tree (Olea europaea L.) genome. *Molecular Genetics and Genomics*, 292(1):53–61.

Barghini, E., Natali, L., Cossu, R. M., Giordani, T., Pindo, M., Cattonaro, F., et al. (2014). The Peculiar Landscape of Repetitive Sequences in the Olive (Olea europaea L.) Genome. *Genome Biology and Evolution*, 6(4):776–791.

Barghini, E., Natali, L., Giordani, T., Cossu, R. M., Scalabrin, S., Cattonaro, F., et al. (2015). LTR retrotransposon dynamics in the evolution of the olive (Olea europaea) genome. *DNA Research*, 22(1):91–100.

Barker, M. S., Arrigo, N., Baniaga, A. E., Li, Z., and Levin, D. A. (2016). On the relative abundance of autopolyploids and allopolyploids.

Barker, M. S., Kane, N. C., Matvienko, M., Kozik, A., Michelmore, R. W., Knapp, S. J., and Rieseberg, L. H. (2008). Multiple paleopolyploidizations during the evolution of the compositae reveal parallel patterns of duplicate gene retention after millions of years. *Molecular Biology and Evolution*, 25(11):2445–2455.

Barker, M. S., Vogel, H., and Schranz, M. E. (2009). Paleopolyploidy in the Brassicales: analyses of the Cleome transcriptome elucidate the history of genome duplications in Arabidopsis and other Brassicales. *Genome biology and evolution*, 1(0):391–9.

Bartolini, G., Messeri, C., Prevost, G., Lavee, S., and Klein, I. (1994). Olive tree germplasm: descriptor lists of cultivated varieties in the world. *Acta Horticulturae*, 356(356):116–118.

Baruca Arbeiter, A., Jakše, J., and Bandelj, D. (2014). Paternity analysis of the olive variety "Istrska Belica" and identification of pollen donors by microsatellite markers. *Scientific World Journal*, 2014:208590.

Bazakos, C., Manioudaki, M. E., Sarropoulou, E., Spano, T., and Kalaitzis, P. (2015). 454 pyrosequencing of olive (Olea europaea L.) transcriptome in response to salinity. *PLoS ONE*, 10(11):e0143000.

Bazakos, C., Manioudaki, M. E., Therios, I., Voyiatzis, D., Kafetzopoulos, D., Awada, T., and Kalaitzis, P. (2012). Comparative transcriptome analysis of two olive cultivars in response to NaCl-stress. *PloS one*, 7(8):e42931.

Bedini, G., Garbari, F., and Peruzzi, L. (2010). Chrobase.it CNs for the Italian flora.

Beghé, D., Piotti, A., Satovic, Z., De La Rosa, R., and Belaj, A. (2017). Pollen-mediated gene flow and fine-scale spatial genetic structure in Olea europaea subsp. Europaea var. sylvestris. *Annals of Botany*, 119(4):671–679.

Belaj, A., Cipriani, G., Testolin, R., Rallo, L., and Trujillo, I. (2004a). Characterization and identification of the main Spanish and Italian olive cultivars by simple-sequence-repeat markers. *HortScience*, 39(7):1557–1561.

Belaj, A., Muñoz-Diez, C., Baldoni, L., Satovic, Z., and Barranco, D. (2010). Genetic diversity and relationships of wild and cultivated olives at regional level in Spain. *Scientia Horticulturae*, 124(3):323–330.

Belaj, A., Satovic, Z., Cipriani, G., Baldoni, L., Testolin, R., Rallo, L., and Trujillo, I. (2003). Comparative study of the discriminating capacity of RAPD, AFLP and SSR markers and of their effectiveness in establishing genetic relationships in olive. *Theoretical and Applied Genetics*, 107(4):736–744.

Belaj, A., Satovic, Z., Trujillo, I., and Rallo, L. (2004b). Genetic relationships of Spanish olive cultivars using RAPD markers. *HortScience*, 39(5):948–951.

Belaj, A., Trujillo, I., de la Rosa, R., Rallo, L., and Gimenez, M. J. (2001). Polymorphism and discrimination capacity of randomly amplified polymorphic markers in an olive germplasm bank. *Journal of the American Society for Horticultural Science*, 126(1):64–71.

Bernhardt, N., Brassac, J., Kilian, B., and Blattner, F. R. (2017). Dated tribe-wide whole chloroplast genome phylogeny indicates recurrent hybridizations within Triticeae. *BMC Evolutionary Biology*, 17(1):1–16.

Besnard, G. (2008). Chloroplast DNA variations in Mediterranean olive. *Journal of Horticultural Science and Biotechnology*, 83(1):51–54.

Besnard, G. and Baali-Cherif, D. (2009). Coexistence of diploids and triploids in a Saharan relict olive: Evidence from nuclear microsatellite and flow cytometry analyses. *Comptes Rendus - Biologies*, 332(12):1115–1120.

Besnard, G., Bakkali, A. E., Haouane, H., Baali-Cherif, D., Moukhli, A., and Khadari, B. (2013a). Population genetics of Mediterranean and Saharan olives: Geographic patterns of differentiation and evidence for early generations of admixture. *Annals of Botany*, 112(7):1293–1302.

Besnard, G., Baradat, P., and Bervillé, A. (2001a). Genetic relationships in the olive (Olea europaea L.) reflect multilocal selection of cultivars. *Theoretical and Applied Genetics*, 102(2-3):251–258.

Besnard, G., Baradat, P., Chevalier, D., Tagmount, A., and Bervillé, A. (2001b). Genetic differentiation in the olive complex (Olea europaea) revealed by RAPDs and RFLPs in the rRNA genes. *Genetic Resources and Crop Evolution*, 48(2):165–182.

Besnard, G. and Bervillé, A. (2000). Multiple origins for Mediterranean olive (Olea europaea L-ssp europaea) based upon mitochondrial DNA polymorphisms. *Comptes Rendus De L Academie Des Sciences Serie Iii-Sciences De La Vie-Life Sciences*, 323(2):173–181.

Besnard, G. and Bervillé, A. (2002). On chloroplast DNA variations in the olive (Olea europaea L.) complex: Comparison of RFLP and PCR polymorphisms. *Theoretical and Applied Genetics*, 104(6-7):1157–1163.

Besnard, G., Dupuy, J., Larter, M., Cuneo, P., Cooke, D., and Chikhi, L. (2014). History of the invasive African olive tree in Australia and Hawaii: Evidence for sequential bottlenecks and hybridization with the Mediterranean olive. *Evolutionary Applications*, 7(2):195–211.

Besnard, G. and El Bakkali, A. (2014). Sequence analysis of single-copy genes in two wild olive subspecies: nucleotide diversity and potential use for testing admixture. *Genome / National Research Council Canada = Génome / Conseil national de recherches Canada*, 57(3):145–53.

Besnard, G., Garcia-Verdugo, C., De Casas, R. R., Treier, U. A., Galland, N., and Vargas, P. (2008). Polyploidy in the olive complex (Olea europaea): evidence from flow cytometry and nuclear microsatellite analyses. *Annals of botany*, 101(1):25–30.

Besnard, G., Henry, P., Wille, L., Cooke, D., and Chapuis, E. (2007a). On the origin of the invasive olives (Olea europaea L., Oleaceae). *Heredity*, 99(6):608–619.

Besnard, G., Hernández, P., Khadari, B., Dorado, G., and Savolainen, V. (2011). Genomic profiling of plastid DNA variation in the Mediterranean olive tree. *BMC plant biology*, 11(1):80.

Besnard, G., Khadari, B., Baradat, P., and Bervillé, A. (2002a). Combination of chloroplast and mitochondrial DNA polymorphisms to study cytoplasm genetic differentiation in the olive complex (Olea europaea L.). *Theoretical and Applied Genetics*, 105(1):139–144.

Besnard, G., Khadari, B., Baradat, P., and Bervillé, A. (2002b). Olea europaea (Oleaceae) phylogeography based on chloroplast DNA polymorphism. *Theoretical and Applied Genetics*, 104(8):1353–1361.

Besnard, G., Khadari, B., Navascués, M., Fernández-Mazuecos, M., El Bakkali, A., Arrigo, N., et al. (2013b). The complex history of the olive tree: from Late Quaternary diversification of Mediterranean lineages to primary domestication in the northern Levant. *Proceedings. Biological sciences / The Royal Society*, 280(1756):20122833.

Besnard, G., Khadari, B., Villemur, P., and Bervillé, A. (2000). Cytoplasmic male sterility in the olive (Olea europaea L.). *Theoretical and Applied Genetics*, 100(7):1018–1024.

Besnard, G. and Rubio de Casas, R. (2016). Single vs multiple independent olive domestications: The jury is (still) out.

Besnard, G., Rubio De Casas, R., Christin, P. A., and Vargas, P. (2009). Phylogenetics of Olea (Oleaceae) based on plastid and nuclear ribosomal DNA sequences: Tertiary climatic shifts and lineage differentiation times. *Annals of Botany*, 104(1):143–160.

Besnard, G., Rubio De Casas, R., and Vargas, P. (2007b). Plastid and nuclear DNA polymorphism reveals historical processes of isolation and reticulation in the olive tree complex (Olea europaea). *Journal of Biogeography*, 34(4):736–752.

Besnard, L. R. G., Martin, A. B. A., la Rosa, R., Angiolillo, A., Guerrero, C., Pellegrini, M., et al. (2003). A first linkage map of olive (Olea europaea L.) cultivars using RAPD, AFLP, RFLP and SSR markers. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, 106(January):1273–1282.

Birol, I., Raymond, A., Jackman, S. D., Pleasance, S., Coope, R., Taylor, G. A., et al. (2013). Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. *Bioinformatics (Oxford, England)*, 29(12):1492–7.

Bitonti, M. B., Cozza, R., Chiappetta, A., Contento, A., Minelli, S., Ceccarelli, M., et al. (1999). Amount and organization of the heterochromatin in Olea europaea and related species. *Heredity*, 83 ( Pt 2):188–95.

Blanc, G. (2004). Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes. *The Plant Cell Online*, 16(7):1667–1678.

Blanc, G., Hokamp, K., and Wolfe, K. H. (2003). A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome Research*, 13(2):137–144.

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics (Oxford, England)*, 27(4):578–9.

Boetzer, M. and Pirovano, W. (2012). Toward almost closed genomes with GapFiller. *Genome biology*, 13(6):R56.

Boetzer, M. and Pirovano, W. (2014). SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC bioinformatics*, 15:211.

Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., et al. (2012). Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics (Oxford, England)*, 28(3):423–5.

Bomblies, K. and Madlung, A. (2014). Polyploidy in the Arabidopsis genus.

Borodovsky, M. and Lomsadze, A. (2011). Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, Chapter 4:Unit 4.6.1–10.

Bosso, L., Di Febbraro, M., Cristinzio, G., Zoina, A., and Russo, D. (2016). Shedding light on the effects of climate change on the potential distribution of Xylella fastidiosa in the Mediterranean basin. *Biological Invasions*, 18(6):1759–1768.

Bowers, J. E., Chapman, B. A., Rong, J., and Paterson, A. H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, 422(6930):433–8.

Bracci, T., Busconi, M., Fogher, C., and Sebastiani, L. (2011). Molecular studies in olive (Olea europaea L.): Overview on DNA markers applications and recent advances in genome analysis.

Bracci, T., Sebastiani, L., Busconi, M., Fogher, C., Belaj, A., and Trujillo, I. (2009). SSR markers reveal the uniqueness of olive cultivars from the Italian region of Liguria. *Scientia Horticulturae*, 122(2):209–215.

Brandizzi, F. and Grilli Caiola, M. (1998). Flow cytometric analysis of nuclear DNA inCrocus sativus and allies (Iridaceae). *Plant Systematics and Evolution*, 211(3-4):149–154.

Breton, C., Terral, J. F., Pinatel, C., Médail, F., Bonhomme, F., and Bervillé, A. (2009). The origins of the domestication of the olive tree. *Comptes Rendus - Biologies*, 332(12):1059–1064.

Breton, C., Tersac, M., and Bervillé, A. (2006). Genetic diversity and gene flow between the wild olive (oleaster, Olea europaea L.) and the olive: several PlioPleistocene refuge zones in the Mediterranean basin suggested by simple sequence repeats analysis. *Journal of Biogeography*, 33(11):1916–1928.

Breton, C. M., Villemur, P., and Bervillé, A. J. (2017). The sporophytic self-incompatibility mating system is conserved in Olea europaea subsp. cuspidata and O. e. europaea. *Euphytica*, 213(1):22.

Breuert, S., Allers, T., Spohn, G., and Soppa, J. (2006). Regulated polyploidy in halophilic archaea. *PLoS ONE*, 1(1):e92.

Brito, G., Loureiro, J., Lopes, T., Rodriguez, E., and Santos, C. (2007). Genetic characterisation of olive trees from Madeira Archipelago using flow cytometry and microsatellite markers. *Genetic Resources and Crop Evolution*, 55(5):657–664.

Bronzini de Caraffa, V., Maury, J., Gambotti, C., Breton, C., Bervillé, A., and Giannettini, J. (2002). Mitochondrial DNA variation and RAPD mark oleasters, olive and feral olive from Western and Eastern Mediterranean. *Theoretical and Applied Genetics*, 104(6-7):1209–1216.

Brummer, E. C., Cazcarro, P. M., and Luth, D. (1999). Ploidy determination of alfalfa germplasm accessions using flow cytometry. *Crop Science*, 39(4):1202–1207.

Bryant, D. and Moulton, V. (2004). Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. *Molecular Biology and Evolution*, 21(2):255–265.

Buerkle, C. A. (2005). Maximum-likelihood estimation of a hybrid index based on molecular markers. *Molecular Ecology Notes*, 5(3):684–687.

Buerkle, C. A., Morris, R. J., Asmussen, M. A., and Rieseberg, L. H. (2000). The likelihood of homoploid hybrid speciation. *Heredity*, 84(4):441–451.

Burton, R. S., Pereira, R. J., and Barreto, F. S. (2013). Cytonuclear Genomic Interactions and Hybrid Breakdown. *Annual Review of Ecology, Evolution, and Systematics*, 44(1):281–302.

Cabezas, J. A., Ibáñez, J., Lijavetzky, D., Vélez, D., Bravo, G., Rodríguez, V., et al. (2011). A 48 SNP set for grapevine cultivar identification. *BMC Plant Biology*, 11(1):153.

Cacciola, S. O., Faedda, R., Sinatra, F., Agosteo, G. E., Schena, L., Frisullo, S., and Magnano di San Lio, G. (2012). Olive anthracnose. *Journal of Plant Pathology*, 94(1):29–44.

Cáceres, M. E., Ceccarelli, M., Pupilli, F., Sarri, V., and Mencuccini, M. (2015). Obtainment of inter-subspecific hybrids in olive (Olea europaea L.). *Euphytica*, 201(2):307–319.

Caceres, M. E., Pupilli, F., Sarri, V., Mencuccini, M., and Ceccarelli, M. (2016). Floral biology in Olea europaea subsp. cuspidata: A comparative structural and functional characterization. *Flora - Morphology, Distribution, Functional Ecology of Plants*, 222:27–36.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST+: architecture and applications. *BMC bioinformatics*, 10(1):421.

Campbell, M. A., Ganley, A. R. D., Gabaldón, T., and Cox, M. P. (2016). The Case of the Missing Ancient Fungal Polyploids. *The American Naturalist*, 188(6):602–614.

Cannon, S. B., McKain, M. R., Harkess, A., Nelson, M. N., Dash, S., Deyholos, M. K., et al. (2015). Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. *Molecular Biology and Evolution*, 32(1):193–210.

Cannon, S. B., Sterck, L., Rombauts, S., Sato, S., Cheung, F., Gouzy, J., et al. (2006). Legume genome evolution viewed through the Medicago truncatula and Lotus japonicus genomes. *Proceedings of the National Academy of Sciences*, 103(40):14959–14964.

Cao, Y., Tian, L., Gao, Y., and Liu, F. (2012). Genetic diversity of cultivated and wild Ussurian Pear (Pyrus ussuriensis Maxim.) in China evaluated with M13-tailed SSR markers. *Genetic Resources and Crop Evolution*, 59(1):9–17.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)*, 25(15):1972–3.

Cardona, G., Rosselló, F., and Valiente, G. (2008). A perl package and an alignment tool for phylogenetic networks. *BMC bioinformatics*, 9(1):175.

Carmona, R., Zafra, A., Seoane, P., Castro, A. J., Guerrero-Fernández, D., Castillo-Castillo, T., et al. (2015). ReprOlive: a database with linked data for the olive tree (Olea europaea L.) reproductive transcriptome. *Frontiers in Plant Science*, 6:625.

Castaings, L., Camargo, A., Pocholle, D., Gaudon, V., Texier, Y., Boutet-Mercey, S., et al. (2009). The nodule inception-like protein 7 modulates nitrate sensing and metabolism in Arabidopsis. *Plant Journal*, 57(3):426–435.

Castelli, M., Miller, C. H., and Schmidt-Lebuhn, A. N. (2017). Polyploidization and Genome Size Evolution in Australian Billy Buttons ( Craspedia , Asteraceae: Gnaphalieae). *International Journal of Plant Sciences*, 178(5):352–361.

Chagné, D., Kirk, C., Whitworth, C., Erasmuson, S., Bicknell, R., Sargent, D. J., et al. (2015). Polyploid and aneuploid detection in apple using a single nucleotide polymorphism array. *Tree Genetics and Genomes*, 11(5):94.

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A. P., Tang, H., Wang, X., et al. (2014). Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science*, 345(6199):950–953.

Chantret, N., Salse, J., Sabot, F., Rahman, S., Bellec, A., Laubin, B., et al. (2005). Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (Triticum and Aegilops). *The Plant Cell*, 17(4):1033–1045.

Chapman, M. A. and Burke, J. M. (2007). Genetic divergence and hybrid speciation. *Evolution*, 61(7):1773–1780.

Chase, M. W., Christenhusz, M. J., Fay, M. F., Byng, J. W., Judd, W. S., Soltis, D. E., et al. (2016). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society*, 181(1):1–20.

Chen, M., Wang, F., Zhang, Z., Fu, J., and Ma, Y. (2017). Characterization of fungi resistance in two autotetraploid apple cultivars. *Scientia Horticulturae*, 220:27–35.

Chen, Z. J. (2007). Genetic and Epigenetic Mechanisms for Gene Expression and Phenotypic Variation in Plant Polyploids. *Annual Review of Plant Biology*, 58(1):377–406.

Chen, Z. J. (2010). Molecular mechanisms of polyploidy and hybrid vigor. *Trends in Plant Science*, 15(2):57–71.

Chen, Z. J. and Ni, Z. (2006). Mechanisms of genomic rearrangements and gene expression changes in plant polyploids.

Chopra, V. L. and Swaminathan, M. S. (1960). Induction of polyploidy in watermelon. *Proceedings of the Indian Academy of Sciences - Section B*, 51(2):57–65.

Chu, Y., Su, X., Huang, Q., and Zhang, X. (2009). Patterns of DNA sequence variation at candidate gene loci in black poplar (Populus nigra L.) as revealed by single nucleotide polymorphisms. *Genetica*, 137(2):141–150.

Cicatelli, A., Fortunati, T., De Feis, I., and Castiglione, S. (2013). Oil composition and genetic biodiversity of ancient and new olive (Olea europea L.) varieties and accessions of southern Italy. *Plant Science*, 210:82–92.

Ciccarese, F., Ambrico, A., Longo, O., and Schiavone, D. (2002). Search for resistance to Verticillium-Wilt and leaf spot in Olive. *Acta Horticulturae*, 586:717–720.

Cipriani, G., Marrazzo, M. T., Marconi, R., Cimato, A., and Testolin, R. (2002). Microsatellite markers isolated in olive (Olea europaea L.) are suitable for individual fingerprinting and reveal polymorphism within ancient cultivars. *Theoretical and Applied Genetics*, 104(2-3):223–228.

Clark, L. V. and Schreier, A. D. (2017). Resolving microsatellite genotype ambiguity in populations of allopolyploid and diploidized autopolyploid organisms using negative correlations between allelic variables.

Claros, M. G., Crespillo, R., Aguilar, M. L., and Cánovas, F. M. (2000). DNA fingerprinting and classification of geographically related genotypes of olive-tree (Olea europaea L.). *Euphytica*, 116(2):131–142.

Comeron, J. M. (1995). A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *Journal of Molecular Evolution*, 41(6):1152–1159.

Consolandi, C., Palmieri, L., Doveri, S., Maestri, E., Marmiroli, N., Reale, S., et al. (2007). Olive variety identification by ligation detection reaction in a universal array format. *Journal of Biotechnology*, 129(3):565–574.

Contento, A., Ceccarelli, M., Gelati, M. T., Maggini, F., Baldoni, L., and Cionini, P. G. (2002). Diversity of Olea genotypes and the origin of cultivated olives. *Theoretical and Applied Genetics*, 104(8):1229–1238.

Corrêa dos Santos, R. A., Goldman, G. H., and Riaño-Pachón, D. M. (2017). ploidyNGS: visually exploring ploidy with Next Generation Sequencing data. *Bioinformatics*, 30(16):2576–2583.

Corrochano, L. M., Kuo, A., Marcet-Houben, M., Polaino, S., Salamov, A., Villalobos-Escobedo, J. M., et al. (2016). Expansion of Signal Transduction Pathways in Fungi by Extensive Genome Duplication. *Current Biology*, 26(12):1577–1584.

Cronn, R. C., Small, R. L., Haselkorn, T., and Wendel, J. F. (2002). Rapid diversification of the cotton genus (Gossypium: Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. *American Journal of Botany*, 89(4):707–725.

Cruz, F., Julca, I., Gómez-Garrido, J., Loska, D., Marcet-Houben, M., Cano, E., et al. (2016a). Genome sequence of the olive tree, Olea europaea. *GigaScience*, 5(1):1–12.

Cruz, F., Julca, I., Gómez-Garrido, J., Loska, D., Marcet-Houben, M., Cano, E., et al. (2016b). Genomics data from the Mediterranean olive tree. GigaScience Database: Olea europaea var. europaea.

Cruz, F., Julca, I., Gómez-Garrido, J., Loska, D., Marcet-Houben, M., Cano, E., et al. (2016c). Olive genome and annotation files.

Culley, T. M. and Hardiman, N. A. (2009). The role of intraspecific hybridization in the evolution of invasiveness: a case study of the ornamental pear tree Pyrus calleryana. *Biological Invasions*, 11(5):1107–1119.

Cutcutache, I., Wu, A. Y., Suzuki, Y., McPherson, J. R., Lei, Z., Deng, N., et al. (2016). Abundant copy-number loss of CYCLOPS and STOP genes in gastric adenocarcinoma. *Gastric Cancer*, 19(2):453–465.

Dai, X., Hu, Q., Cai, Q., Feng, K., Ye, N., Tuskan, G. A., et al. (2014). The willow genome and divergent evolution from poplar after the common genome duplication. *Cell Research*, 24(10):1274–1277.

Dart, S., Kron, P., and Mable, B. K. (2004). Characterizing polyploidy in Arabidopsis lyrata using chromosome counts and flow cytometry. *Canadian Journal of Botany-Revue Canadienne De Botanique*, 82(2):185–197.

Darwin, C. (1859). On the origin of the species by natural selection.

Das, A. B., Rai, S., and Das, P. (1998). Estimation of 4C DNA and Karyotype Analysis in Ginger (Zingiber officinale Rosc)II. *Cytologia*, 63(2):133–139.

De Smet, R., Adams, K. L., Vandepoele, K., Van Montagu, M. C. E., Maere, S., and Van de Peer, Y. (2013). Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences of the United States of America*, 110(8):2898–903.

Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., et al. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science (New York, N.Y.)*, 345(6201):1181–4.

D'Hont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, F., Garsmeur, O., et al. (2012). The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature*, 488(7410):213–217.

Diamond, J. (2002). Evolution, consequences and future of plant and animal domestication. *Nature*, 418(6898):700–707.

Díaz, A. (2012). Olive. In *Technological Innovations in Major World Oil Crops, Volume 1*, pages 267–291. Springer New York, New York, NY.

Díaz, A., De La Rosa, R., Martín, A., and Rallo, P. (2006). Development, characterization and inheritance of new microsatellites in olive (Olea europaea L.) and evaluation of their usefulness in cultivar identification and genetic relationship studies. *Tree Genetics and Genomes*, 2(3):165–175.

Díaz, A., Martín-Hernández, A. M., Dolcet-Sanjuan, R., Garcés-Claver, A., Álvarez, J. M., Garcia-Mas, J., et al. (2017). Quantitative trait loci analysis of melon (Cucumis melo L.) domestication-related traits.

Díaz, A., Zarouri, B., Fergany, M., Eduardo, I., Álvarez, J. M., Picó, B., and Monforte, A. J. (2014). Mapping and introgression of QTL involved in fruit shape transgressive segregation into 'Piel de Sapo' melon (Cucucumis melo L.). *PLoS ONE*, 9(8):e104188.

Díez, C. M. and Gaut, B. S. (2016). The jury may be out, but it is important that it deliberates: A response to Besnard and Rubio de Casas about olive domestication.

Diez, C. M., Trujillo, I., Martinez-Urdiroz, N., Barranco, D., Rallo, L., Marfil, P., and Gaut, B. S. (2015). Olive domestication and diversification in the Mediterranean Basin. *New Phytologist*, 206(1):436–447.

Dighton, A., Fairbairn, A., Bourke, S., Faith, J. T., and Habgood, P. (2017). Bronze Age olive domestication in the north Jordan valley: new morphological evidence for regional complexity in early arboricultural practice from Pella in Jordan. *Vegetation History and Archaeobotany*, 26(4):403–413.

Divakar, P. K., Del-Prado, R., Thorsten Lumbsch, H., Wedin, M., Esslinger, T. L., Leavitt, S. D., and Crespo, A. (2012). Diversification of the newly recognized lichen-forming fungal lineage Montanelia (Parmeliaceae, Ascomycota) and its relation to key geological and climatic events. *American Journal of Botany*, 99(12):2014–2026.

Dodsworth, S., Chase, M. W., and Leitch, A. R. (2016). Is post-polyploidization diploidization the key to the evolutionary success of angiosperms? *Botanical Journal of the Linnean Society*, 180(1):1–5.

Doebley, J. F., Gaut, B. S., and Smith, B. D. (2006). The Molecular Genetics of Crop Domestication.

Doležel, J., Greilhuber, J., and Suda, J. (2007). Estimation of nuclear DNA content in plants using flow cytometry. *Nature Protocols*, 2(9):2233–2244.

Donini, P., Sarri, V., Baldoni, L., Porceddu, A., Cultrera, N., Contento, A., et al. (2006). Microsatellite markers are powerful tools for discriminating among olive cultivars and assigning them to geographically defined populations. *Genome*, 49(12):1606–1615.

Doungous Oumar, D., Sama, A., Adiobo, A., and Zok, S. (2011). Determination of ploidy level by flow cytometry and autopolyploid induction in cocoyam (Xanthosoma sagittifolium). *African Journal of Biotechnology*, 1O(73):16491–16494.

Doyle, J. J. and Sherman-Broyles, S. (2017). Double trouble: taxonomy and definitions of polyploidy.

Doyon, J. P., Scornavacca, C., Gorbunov, K. Y., Szöllosi, G. J., Ranwez, V., and Berry, V. (2010). An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6398 LNBI, pages 93–108. Springer, Berlin, Heidelberg.

Drescher, A., Stephanie, R., Calsa, T., Carrer, H., and Bock, R. (2000). The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. *Plant Journal*, 22(2):97–104.

Dufresne, F., Stift, M., Vergilino, R., and Mable, B. K. (2014). Recent progress and challenges in population genetics of polyploid organisms: An overview of current state-of-the-art molecular and statistical tools.

Duran, C., Edwards, D., and Batley, J. (2009). Genetic maps and the use of synteny. *Methods in Molecular Biology*, 513:41–55.

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

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–7.

El Aabidine, A. Z., Charafi, J., Grout, C., Doligez, A., Santoni, S., Moukhli, A., et al. (2010). Construction of a genetic linkage map for the olive based on AFLP and SSR markers. *Crop Science*, 50(6):2291–2302.

El Baidouri, M., Murat, F., Veyssiere, M., Molinier, M., Flores, R., Burlot, L., et al. (2017). Reconciling the evolutionary origin of bread wheat (Triticum aestivum). *New Phytologist*, 213(3):1477–1486.

Ellis, C. M., Nagpal, P., Young, J. C., Hagen, G., Guilfoyle, T. J., and Reed, J. W. (2005). AUXIN RESPONSE FACTOR1 and AUXIN RESPONSE FACTOR2 regulate senescence and floral organ abscission in Arabidopsis thaliana. *Development*, 132(20):4563–4574.

Ellstrand, N. C. and Schierenbeck, K. A. (2006). Hybridization as a stimulus for the evolution of invasiveness in plants? *Euphytica*, 148(1-2):35–46.

Eric Schranz, M., Mohammadin, S., and Edger, P. P. (2012). Ancient whole genome duplications, novelty and diversification: The WGD Radiation Lag-Time Model.

Erickson, D. L., Fenster, C. B., Rickson, D. A. L. E., and Enster, C. H. B. F. (2006). Intraspecific Hybridization and the Recovery of Fitness in the Native Legume Chamaecrista Fasciculata. *Evolution*, 60(2):225–233.

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

Fabbri, A., Hormaza, J. I., and Polito, V. S. (1995). Random Amplified Polymorphic DNA Analysis of Olive ( Olea europaea L .) Cultivars. *Journal of the American Society for Horticultural Science*, 120(3):538–542.

Falistocco, E. and Tosti, N. (1996). Cytogenetic investigation in Olea europea L. *Journal of Genetics & Breeding*, 50(3):235–238.

Fang, Z. and Morrell, P. L. (2016). Domestication: Polyploidy boosts domestication. *Nature Plants*, 2(8):16116.

FAOSTAT (2014). Statistical database. http://www.fao.org/faostat/en/{#}data/QC.

Fawcett, J. A., Maere, S., and Van de Peer, Y. (2009). Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proceedings of the National Academy of Sciences of the United States of America*, 106(14):5737–42.

Fehrer, J., Gemeinholzer, B., Chrtek, J., and Bräutigam, S. (2007). Incongruent plastid and nuclear DNA phylogenies reveal ancient intergeneric hybridization in Pilosella hawkweeds (Hieracium, Cichorieae, Asteraceae). *Molecular Phylogenetics and Evolution*, 42(2):347–361.

Feuillet, C., Travella, S., Stein, N., Albar, L., Nublat, A., and Keller, B. (2003). Map-based isolation of the leaf rust disease resistance gene Lr10 from the hexaploid wheat (Triticum aestivum L.) genome. *Proceedings of the National Academy of Sciences of the United States of America*, 100(25):15253–15258.

Figueras, A., Robledo, D., Corvelo, A., Hermida, M., Pereiro, P., Rubiolo, J. A., et al. (2016). Whole genome sequencing of turbot (Scophthalmus maximus; Pleuronectiformes): A fish adapted to demersal life. *DNA Research*, 23(3):181–192.

Fishman, L., Willis, J. H., Wu, C. A., and Lee, Y.-W. (2014). Comparative linkage maps suggest that fission, not polyploidy, underlies near-doubling of chromosome number within monkeyflowers (Mimulus; Phrymaceae). *Heredity*, 112(5):562–8.

Fleischmann, A., Michael, T. P., Rivadavia, F., Sousa, A., Wang, W., Temsch, E. M., et al. (2014). Evolution of genome size and chromosome number in the carnivorous plant genus Genlisea (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms. *Annals of Botany*, 114(8):1651–1663.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–1545.

Freyman, W. A. and Höhna, S. (2017). Cladogenetic and Anagenetic Models of Chromosome Number Evolution: A Bayesian Model Averaging Approach. *Systematic Biology*, 103:1252–1258.

Frias, L. and Ribeca, P. (2016). ASM scripts. https://github.com/lfrias81/anchor-asm/tree/master/wrapper.

Galla, G., Barcaccia, G., Ramina, A., Collani, S., Alagna, F., Baldoni, L., et al. (2009). Computational annotation of genes differentially expressed along olive fruit development. *BMC plant biology*, 9(1):128.

Gallagher, J. P., Grover, C. E., Hu, G., and Wendel, J. F. (2016). Insights into the Ecology and Evolution of Polyploid Plants through Network Analysis. *Molecular Ecology*, 25(11):2644–2660.

Garcia-Mas, J., Benjak, A., Sanseverino, W., Bourgeois, M., Mir, G., González, V. M., et al. (2012). The genome of melon (Cucumis melo L.). *Proceedings of the National Academy of Sciences of the United States of America*, 109(29):11872–7.

García-Verdugo, C., Fay, M. F., Granado-Yela, C., DE Casas, R. R., Balaguer, L., Besnard, G., and Vargas, P. (2009). Genetic diversity and differentiation processes in the ploidy series of Olea europaea L.: a multiscale approach from subspecies to insular populations. *Molecular ecology*, 18(3):454–67.

Garsmeur, O., Charron, C., Bocs, S., Jouffe, V., Samain, S., Couloux, A., et al. (2011). High homologous gene conservation despite extreme autopolyploid redundancy in sugarcane. *New Phytologist*, 189(2):629–642.

Gebhardt, C., Walkemeier, B., Henselewski, H., Barakat, A., Delseny, M., and Stüber, K. (2003). Comparative mapping between potato (Solanum tuberosum) and Arabidopsis thaliana reveals structurally conserved domains and ancient duplications in the potato genome. *The Plant journal : for cell and molecular biology*, 34(4):529–41.

Giampetruzzi, A., Saponari, M., Almeida, R. P. P., Essakhi, S., Boscia, D., Loconsole, G., and Saldarelli, P. (2017). Complete Genome Sequence of the Olive-Infecting Strain Xylella fastidiosa subsp. pauca De Donno. *Genome announcements*, 5(27):e00569–17.

Gill, N., Findley, S., Walling, J. G., Hans, C., Ma, J., Doyle, J., et al. (2009). Molecular and chromosomal evidence for allopolyploidy in soybean. *Plant physiology*, 151(3):1167–74.

Glick, L. and Mayrose, I. (2014). ChromEvol: Assessing the pattern of chromosome number evolution and the inference of polyploidy along a phylogeny. *Molecular Biology and Evolution*, 31(7):1914–1922.

Glover, N. M., Redestig, H., and Dessimoz, C. (2016). Homoeologs: What Are They and How Do We Infer Them?

Goff, S. A. (2002). A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. japonica). *Science*, 296(5565):92–100.

Goldblatt, P. (1979). Polyploidy in angiosperms: monocotyledons. *Basic Life Science*, 13:219–239.

Goldblatt, P. and Lowry, P. P. (2011). The Index to Plant Chromosome Numbers (Ipcn): Three Decades of Publication by the Missouri Botanical Garden Come to An End. *Annals of the Missouri Botanical Garden*, 98(2):226–227.

Gompert, Z. and Alex Buerkle, C. (2010). Introgress: A software package for mapping components of isolation in hybrids. *Molecular Ecology Resources*, 10(2):378–384.

González-Plaza, J. J., Ortiz-Martín, I., Muñoz-Mérida, A., García-López, C., Sánchez-Sevilla, J. F., Luque, F., et al. (2016). Transcriptomic Analysis Using Olive Varieties and Breeding Progenies Identifies Candidate Genes Involved in Plant Architecture. *Frontiers in plant science*, 7:240.

Goodman, M., Czelusniak, J., Moore, G. W., Romero-Herrera, A. E., and Matsuda, G. (1979). Fitting the Gene Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences. *Systematic Biology*, 28(2):132–163.

Gostinčar, C., Ohm, R. A., Kogej, T., Sonjak, S., Turk, M., Zajc, J., et al. (2014). Genome sequencing of four Aureobasidium pullulans varieties: biotechnological potential, stress tolerance, and description of new species. *BMC genomics*, 15:549.

Goulet, B. E., Roda, F., and Hopkins, R. (2017). Hybridization in Plants: Old Ideas, New Techniques. *Plant Physiology*, 173(1):65–78.

Goulet, C., Mageroy, M. H., Lam, N. B., Floystad, A., Tieman, D. M., and Klee, H. J. (2012). Role of an esterase in flavor volatile variation within the tomato clade. *Proceedings of the National Academy of Sciences*, 109(46):19009–19014.

Grant, V. (1958). The regulation of recombination in plants. *Cold Spring Harbor symposia on quantitative biology*, 23:337–63.

Green, P. (2002). A revision of Olea L. (Oleaceae). *Springer*, 54(1):91–140.

Green, P. and Wickens, G. (1989). *The Olea europaea complex*. The Davis & Hedge Festschrift, Edinburg University Press, Edinburgh.

Green, P. S. (2004). Oleaceae. In *Flowering Plants · Dicotyledons*, pages 296–306. Springer Berlin Heidelberg, Berlin, Heidelberg.

Gregg, W. C. T., Ather, S. H., and Hahn, M. W. (2017). Gene-Tree Reconciliation with MUL-Trees to Resolve Polyploidy Events. *Systematic Biology*, pages 1–12.

Gregory, T. R. (2011). *The Evolution of the Genome*. Elsevier Academic.

Griese, M., Lange, C., and Soppa, J. (2011). Ploidy in cyanobacteria. *FEMS Microbiology Letters*, 323(2):124–131.

Guerra, D., Lamontanara, A., Bagnaresi, P., Orrù, L., Rizza, F., Zelasco, S., et al. (2015). Transcriptome changes associated with cold acclimation in leaves of olive tree (Olea europaea L.). *Tree Genetics and Genomes*, 11(6):113.

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

Guiu-Aragonés, C., Monforte, A. J., Saladié, M., Corrêa, R. X., Garcia-Mas, J., and Martín-Hernández, A. M. (2014). The complex resistance to cucumber mosaic cucumovirus (CMV) in the melon accession PI161375 is governed by one gene and at least two quantitative trait loci. *Molecular Breeding*, 34(2):351–362.

Guo, H. S., Zhang, Y. M., Sun, X. Q., Li, M. M., Hang, Y. Y., and Xue, J. Y. (2016). Evolution of the KCS gene family in plants: the history of gene duplication, sub/neofunctionalization and redundancy. *Molecular genetics and genomics : MGG*, 291(2):739–752.

Guyot, R. and Keller, B. (2004). Ancestral genome duplication in rice. *Genome*, 47(3):610–614.

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K., Hannick, L. I., et al. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic acids research*, 31(19):5654–66.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology*, 9(1):R7.

Hannachi, H., Sommerlatte, H., Breton, C., Msallem, M., El Gazzah, M., Ben El Hadj, S., and Bervillé, A. (2009). Oleaster (var. sylvestris) and subsp. cuspidata are suitable genetic resources for improvement of the olive (Olea europaea subsp. europaea var. europaea). *Genetic Resources and Crop Evolution*, 56(3):393–403.

Havananda, T., Charles Brummer, E., and Doyle, J. J. (2011). Complex patterns of autopolyploid evolution in alfalfa and allies(Medicago Sativa; Leguminosae). *American Journal of Botany*, 98(10):1633–1646.

Hedges, S. B., Marin, J., Suleski, M., Paymer, M., and Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification. *Molecular Biology and Evolution*, 32(4):835–845.

Hellsten, U., Wright, K. M., Jenkins, J., Shu, S., Yuan, Y., Wessler, S. R., et al. (2013). Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 110(48):19478–82.

Hess, J., Kadereit, J. W., and Vargas, P. (2000). The colonization history of Olea europaea L. in Macaronesia based on internal transcribed spacer 1 (ITS-1) sequences, randomly amplified polymorphic DNAs (RAPD), and intersimple sequence repeats (ISSR). *Molecular Ecology*, 9(7):857–868.

Hias, N., Leus, L., Davey, M. W., Vanderzande, S., Van Huylenbroeck, J., and Keulemans, J. (2017). Effect of polyploidization on morphology in two apple (Malus domestica) genotypes. *Horticultural Science*, 44(2):55–63.

Hinsinger, D. D., Basak, J., Gaudeul, M., Cruaud, C., Bertolino, P., Frascaria-Lacoste, N., and 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(11):e80431.

Hinsley, S. (2009). Chromosome counts for Malvaceae.

Hirsch, H., Brunet, J., Zalapa, J. E., von Wehrden, H., Hartmann, M., Kleindienst, C., et al. (2017). Intra- and interspecific hybridization in invasive Siberian elm. *Biological Invasions*, 19(6):1889–1904.

Hobolth, A., Christensen, O. F., Mailund, T., and Schierup, M. H. (2007). Genomic Relationships and Speciation Times of Human, Chimpanzee, and Gorilla Inferred from a Coalescent Hidden Markov Model. *PLoS Genetics*, 3(2):e7.

Hofberger, J. A., Lyons, E., Edger, P. P., Chris Pires, J., and Eric Schranz, M. (2013). Whole genome and tandem duplicate retention facilitated glucosinolate pathway diversification in the mustard family. *Genome Biology and Evolution*, 5(11):2155–2173.

Hosseini-Mazinani, M., Mariotti, R., Torkzaban, B., Sheikh-Hassani, M., Ataei, S., Cultrera, N. G. M., et al. (2014). High genetic diversity detected in olives beyond the boundaries of the Mediterranean sea. *PLoS ONE*, 9(4):e93146.

Huang, J.-L., Sun, G.-L., and Zhang, D.-M. (2010). Molecular evolution and phylogeny of the angiosperm ycf2 gene. *Journal of Systematics and Evolution*, 48(4):240–248.

Huang, X. Y., Chao, D. Y., Gao, J. P., Zhu, M. Z., Shi, M., and Lin, H. X. (2009). A previously unknown zinc finger protein, DST, regulates drought and salt tolerance in rice via stomatal aperture control. *Genes and Development*, 23(15):1805–1817.

Huerta-Cepas, J., Capella-Gutierrez, S., Pryszcz, L. P., Denisov, I., Kormes, D., Marcet-Houben, M., et al. (2011). PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic acids research*, 39(Database issue):D556–60.

Huerta-Cepas, J., Capella-Gutierrez, S., Pryszcz, L. P., Marcet-Houben, M., Gabaldon, T., Capella-Gutiérrez, S., et al. (2014). PhylomeDB v4: Zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Research*, 42(Database issue):D897–902.

Huerta-Cepas, J., Dopazo, H., Dopazo, J., and Gabaldón, T. (2007). The human phylome. *Genome biology*, 8(6):R109.

Huerta-Cepas, J. and Gabaldón, T. (2011). Assigning duplication events to relative temporal scales in genome-wide studies. *Bioinformatics*, 27(1):38–45.

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, 33(6):1635–1638.

Huson, D. H. and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution*, 23(2):254–67.

Huson, D. H. and Scornavacca, C. (2012). Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Systematic Biology*, 61(6):1061–1067.

Hwang, J. H., Seo, D. H., Kang, B. G., Kwak, J. M., and Kim, W. T. (2014). Suppression of Arabidopsis AtPUB30 resulted in increased tolerance to salt stress during germination. *Plant Cell Reports*, 34(2):277–289.

Iaria, D., Chiappetta, A., Muzzalupo, I., Iaria, D., Chiappetta, A., and Muzzalupo, I. (2016). De Novo Transcriptome Sequencing of Olea europaea L. to Identify Genes Involved in the Development of the Pollen Tube. *The Scientific World Journal*, 2016:1–7.

Ibarra-Laclette, E., Lyons, E., Hernández-Guzmán, G., Pérez-Torres, C. A., Carretero-Paulet, L., Chang, T.-H., et al. (2013). Architecture and evolution of a minute plant genome. *Nature*, 498(7452):94–8.

International Peach Genome Initiative, I., Verde, I., Abbott, A. G., Scalabrin, S., Jung, S., Shu, S., et al. (2013). The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nature genetics*, 45(5):487–94.

International Wheat Genome Sequencing Consortium (2014). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science*, 345(6194):1251788–1251788.

Iorizzo, M., Ellison, S., Senalik, D., Zeng, P., Satapoomin, P., Huang, J., et al. (2016). A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nature Genetics*, 48(6):657–666.

İpek, A., İpek, M., Ercişli, S., and Tangu, N. A. (2017). Transcriptome-based SNP discovery by GBS and the construction of a genetic map for olive.

Iwata, H. and Gotoh, O. (2012). Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic acids research*, 40(20):e161.

Jacox, E., Chauve, C., Szöllosi, G. J., Ponty, Y., and Scornavacca, C. (2016). EcceTERA: Comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*, 32(13):2056–2058.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449(7161):463–467.

Jara-Seguel, P. and Urrutia, J. (2011). Chilean plants cytogenetic database.

Jarvis, D. E., Ho, Y. S., Lightfoot, D. J., Schmöckel, S. M., Li, B., Borm, T. J. A., et al. (2017). The genome of Chenopodium quinoa. *Nature*, 542(7641):307–312.

Jeandroz, S., Roy, A., and Bousquet, J. (1997). Phylogeny and phylogeography of the circumpolar genus Fraxinus (Oleaceae) based on internal transcribed spacer sequences of nuclear ribosomal DNA. *Molecular phylogenetics and evolution*, 7(2):241–51.

Jiao, Y., Leebens-Mack, J., Ayyampalayam, S., Bowers, J. E., McKain, M. R., McNeal, J., et al. (2012). A genome triplication associated with early diversification of the core eudicots. *Genome biology*, 13(1):R3.

Jiao, Y., Li, J., Tang, H., and Paterson, A. H. (2014). Integrated Syntenic and Phylogenomic Analyses Reveal an Ancient Genome Duplication in Monocots. *The Plant Cell*, 26(7):2792–2802.

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature*, 473(7345):97–100.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics (Oxford, England)*, 30(9):1236–40.

Julca, I., Marcet-Houben, M., Vargas, P., and Gabaldon, T. (2017). Phylogenomics of the olive tree (Olea europaea) disentangles ancient allo- and autopolyploidizations in Lamiales. *bioRxiv*.

Kang, Y. J., Kim, S. K., Kim, M. Y., Lestari, P., Kim, K. S. K. H., Ha, B.-K., et al. (2014). Genome sequence of mungbean and insights into evolution within Vigna species. *Nature communications*, 5:5443.

Kaniewski, D., Van Campo, E., Boiy, T., Terral, J. F., Khadari, B., and Besnard, G. (2012). Primary domestication and early uses of the emblematic olive tree: Palaeobotanical, historical and molecular evidence from the Middle East. *Biological Reviews*, 87(4):885–899.

Karve, R., Suárez-Román, F., and Iyer-Pascuzzi, A. S. (2016). The Transcription Factor NIN-LIKE PROTEIN7 Controls Border-Like Cell Release. *Plant Physiology*, 171(3):2101–2111.

Katoh, K. and Frith, M. C. (2012). Adding unaligned sequences into an existing alignment using MAFFT and LAST. *Bioinformatics*, 28(23):3144–3146.

Katoh, K., Kuma, K.-i., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic acids research*, 33(2):511–8.

Kaya, H. B., Cetin, O., Kaya, H., Sahin, M., Sefer, F., Kahraman, A., and Tanyolac, B. (2013). SNP Discovery by Illumina-Based Transcriptome Sequencing of the Olive and the Genetic Characterization of Turkish Olive Genotypes Revealed by AFLP, SSR and SNP Markers. *PLoS ONE*, 8(9):e73674.

Kew, R., Gardens, V., and Wakehurst, V. (2012). Plant DNA C-values Database.

Khadari, B., El Aabidine, A. Z., Grout, C., Ben Sadok, I., Doligez, A., Moutier, N., et al. (2010). A Genetic Linkage Map of Olive Based on Amplified Fragment Length Polymorphism, Intersimple Sequence Repeat and Simple Sequence Repeat Markers. *Journal of the American Society for Horticultural Science*, 135(6):548–555.

Khan, A., Belfield, E. J., Harberd, N. P., Mithani, A., Wendel, J. F., Blanc, G., et al. (2016). HANDS2: accurate assignment of homoeallelic base-identity in allopolyploids despite missing data. *Scientific Reports*, 6(July):29234.

Kim, S. T. and Donoghue, M. J. (2008). Incongruence between cpDNA and nrITS trees indicates extensive hybridization within Eupersicaria (Polygonaceae). *American Journal of Botany*, 95(9):1122–1135.

Kimura, M. (1977). Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, 267(May):275–276.

Klepo, T., Toumi, A., De La Rosa, R., León, L., and Belaj, A. (2014). Agronomic evaluation of seedlings from crosses between the main spanish olive cultivar ‘Picual’ and two wild olive trees. *Journal of Horticultural Science and Biotechnology*, 89(5):508–512.

Koenig, D., Jimenez-Gomez, J. M., Kimura, S., Fulop, D., Chitwood, D. H., Headland, L. R., et al. (2013). Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. *Proceedings of the National Academy of Sciences*, 110(28):E2655–E2662.

Kolmogorov, M., Raney, B., Paten, B., and Pham, S. (2014). Ragout–a reference-assisted assembly tool for bacterial genomes. *Bioinformatics*, 30(12):i302–i309.

Kraaijeveld, K. (2010). Genome Size and Species Diversification. *Evolutionary Biology*, 37(4):227–233.

Kürschner, W. M. and Kvaček, Z. (2009). Oligocene-Miocene CO2 fluctuations, climatic and palaeofloristic trends inferred from fossil plant assemblages in central Europe. *Bulletin of Geosciences*, 84(2):189–202.

Lambert, P., Campoy, J. A., Pacheco, I., Mauroux, J. B., Da Silva Linge, C., Micheletti, D., et al. (2016). Identifying SNP markers tightly associated with six major genes in peach [Prunus persica (L.) Batsch] using a high-density SNP array with an objective of marker-assisted selection (MAS). *Tree Genetics and Genomes*, 12(6):121.

Landan, G. and Graur, D. (2007). Heads or tails: A simple reliability check for multiple sequence alignments. *Molecular Biology and Evolution*, 24(6):1380–1383.

Langevin, S., Clay, K., and Grace, J. (1990). The Incidence and Effects of Hybridization between Cultivated Rice and its Related Weed Red Rice (Oryza sativa L.). *Evolution; international journal of organic evolution*, 44(4):1000–1008.

Lassmann, T. and Sonnhammer, E. L. L. (2005). Kalign–an accurate and fast multiple sequence alignment algorithm. *BMC bioinformatics*, 6:298.

Lee, Y. K., Kim, G.-T., Kim, I.-J., Park, J., Kwak, S.-S., Choi, G., and Chung, W.-I. (2006). LONGIFOLIA1 and LONGIFOLIA2, two homologous genes, regulate longitudinal cell elongation in Arabidopsis. *Development*, 133(21):4305–4314.

Leitch, A. R. and Leitch, I. J. (2008). Genomic Plasticity and the Diversity of Polyploid Plants. *Science*, 320(5875):481–483.

Leitch, I. J. and Bennett, M. D. (2004). Genome downsizing in polyploid plants. *Biological Journal of the Linnean Society*, 82(4):651–663.

Leitch, I. J. and Leitch, A. R. (2013). Genome size diversity and evolution in land plants. In *Plant Genome Diversity*, volume 2, pages 307–322. Springer Vienna, Vienna.

Levin, D. A., Francisco-Ortega, J., and Jansen, R. K. (1996). Hybridization and the Extinction of Rare Plant Species. *Source: Conservation Biology*, 10(1):10–16.

Lexer, C., Welch, M. E., Raymond, O., and Rieseberg, L. H. (2003). the Origin of Ecological Divergence in Helianthus Paradoxus (Asteraceae): Selection on Transgressive Characters in a Novel Hybrid Habitat. *Evolution*, 57(9):1989–2000.

Li, B., Cantino, P. D., Olmstead, R. G., Bramley, G. L. C., Xiang, C.-L., Ma, Z.-H., et al. (2016). A large-scale chloroplast phylogeny of the Lamiaceae sheds new light on its subfamilial classification. *Scientific reports*, 6(1):34343.

Li, B., Zhao, Y., Zhu, Q., Zhang, Z., Fan, C., Amanullah, S., et al. (2017). Mapping of powdery mildew resistance genes in melon (Cucumis melo L.) by bulked segregant analysis. *Scientia Horticulturae*, 220:160–167.

Li, D.-Z., Gao, L.-M., Li, H.-T., Wang, H., Ge, X.-J., Liu, J.-Q., et al. (2011). Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences*, 108(49):19641–19646.

Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R. J., et al. (2015a). Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution. *Nature Biotechnology*, 33(5):524–530.

Li, F., Fan, G., Wang, K., Sun, F., Yuan, Y., Song, G., et al. (2014). Genome sequence of the cultivated cotton Gossypium arboreum. *Nature Genetics*, 46(6):567–572.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–60.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Li, J., Alexander, J. H., and Zhang, D. (2002). Paraphyletic Syringa (Oleaceae): Evidence from Sequences of Nuclear Ribosomal DNA ITS and ETS Regions. *Systematic Botany*, 27(3):592–597.

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., et al. (2008). Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science*, 319(5866):1100–1104.

Li, S., Zhao, B., Yuan, D., Duan, M., Qian, Q., Tang, L., et al. (2013). Rice zinc finger protein DST enhances grain production through controlling Gn1a/OsCKX2 expression. *Proceedings of the National Academy of Sciences*, 110(8):3167–3172.

Li, W. H. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution.

Li, Y., Tessaro, M. J., Li, X., and Zhang, Y. (2010). Regulation of the Expression of Plant Resistance Gene SNC1 by a Protein with a Conserved BAT2 Domain. *Plant Physiology*, 153(3):1425–1434.

Li, Y. H., Zhang, W., and Li, Y. (2015b). Transcriptomic analysis of flower blooming in jasminum sambac through De Novo RNA sequencing. *Molecules*, 20(6):10734–10747.

Li, Z., Pinson, S. R. M., Paterson, A. H., Park, W. D., and Stansel, J. W. (1996). Genetics of hybrid sterility and hybrid breakdown in an interspecific rice (Oryza sativa L.) population. *Rice Genetics III - Proceedings of the Third International Rice Genetics Symposium*, 145(1974):409–417.

Lim, P. O., Lee, I. C., Kim, J., Kim, H. J., Ryu, J. S., Woo, H. R., and Nam, H. G. (2010). Auxin response factor 2 (ARF2) plays a major role in regulating auxin-mediated leaf longevity. *Journal of Experimental Botany*, 61(5):1419–1430.

Linder, C. R. and Rieseberg, L. H. (2004). Reconstructing patterns of reticulate evolution in plants.

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., et al. (2013). Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv*, page 1308.2012.

Liu, J., Hua, W., Hu, Z., Yang, H., Zhang, L., Li, R., et al. (2015). Natural variation in ARF18 gene simultaneously affects seed weight and silique length in polyploid rapeseed. *Proceedings of the National Academy of Sciences*, 112(37):E5123–E5132.

Liu, S., Liu, Y., Yang, X., Tong, C., Edwards, D., Parkin, I. A. P., et al. (2014a). The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. *Nature communications*, 5:3930.

Liu, T., Tang, S., Zhu, S., Tang, Q., and Zheng, X. (2014b). Transcriptome comparison reveals the patterns of selection in domesticated and wild ramie (Boehmeria nivea L. Gaud). *Plant Molecular Biology*, 86(1-2):85–92.

Liu, W., Zhao, S., Cheng, Z., Wan, X., Yan, Z., and King, S. (2010). Lycopene and citrulline contents in watermelon (Citrullus lanatus) fruit with different ploidy and changes during fruit development. *Acta Horticulturae*, 871:543–550.

Lloyd, A. and Bomblies, K. (2016). Meiosis in autopolyploid and allopolyploid Arabidopsis.

Logsdon, J., Neiman, M., Boore, J., Sharbrough, J., Bankers, L., McElroy, K., et al. (2017). A very recent whole genome duplication in Potamopyrgus antipodarum predates multiple origins of asexuality & associated polyploidy. *PeerJ Preprints*.

Lohaus, R. and Van de Peer, Y. (2016). Of dups and dinos: Evolution at the K/Pg boundary.

Lopes, M. S., Mendoça, D., Sefc, K. M., Gil, F. S., and Machado, a. D. (2004). Genetic evidence of intra-cutlivar variability within Iberian olive cultivars. *HortScience*, 39(7):1562–1565.

López-Escudero, F. J., Del Río, C., Caballero, J. M., and Blanco-López, M. A. (2004). Evaluation of olive cultivars for resistance to Verticillium dahliae. *European Journal of Plant Pathology*, 110(1):79–85.

López-Girona, E., Zhang, Y., Eduardo, I., Mora, J. R. H., Alexiou, K. G., Arús, P., and Aranzana, M. J. (2017). A deletion affecting an LRR-RLK gene co-segregates with the fruit flat shape trait in peach. *Scientific Reports*, 7(1):6714.

Lott, M., Spillner, A., Huber, K. T., and Moulton, V. (2009). PADRE: A package for analyzing and displaying reticulate evolution. *Bioinformatics*, 25(9):1199–1200.

Loureiro, J., Rodriguez, E., Costa, A., and Santos, C. (2006). Nuclear DNA content estimations in wild olive (Olea europaea L. ssp. europaea var. sylvestris Brot.) and Portuguese cultivars of O. europaea using flow cytometry. *Genetic Resources and Crop Evolution*, 54(1):21–25.

Loureiro, J., Rodriguez, E., Dolezel, J., and Santos, C. (2007). Two new nuclear isolation buffers for plant DNA flow cytometry: a test with 37 species. *Annals of botany*, 100(4):875–88.

Lowe, T. M. and Chan, P. P. (2016). tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic acids research*, 44(W1):W54–W57.

Lumaret, R., Amane, M., Ouazzani, N., Baldoni, L., and Debain, C. (2000). Chloroplast DNA variation in the cultivated and wild olive taxa of the genus Olea L. *Theoretical and Applied Genetics*, 101(4):547–553.

Lumaret, R., Ouazzani, N., Michaud, H., Vivier, G., Deguilloux, M.-F., and Di Giusto, F. (2004). Allozyme variation of oleaster populations (wild olive tree) (Olea europaea L.) in the Mediterranean Basin. *Heredity*, 92(4):343–351.

Lundberg, G., Jin, Y., Sehic, D., Øra, I., Versteeg, R., and Gisselsson, D. (2013). Intratumour Diversity of Chromosome Copy Numbers in Neuroblastoma Mediated by On-Going Chromosome Loss from a Polyploid State. *PLoS ONE*, 8(3):e59268.

Luo, J., Gao, Y., Ma, W., Bi, X.-y., Wang, S.-y., Wang, J., et al. (2014). Tempo and mode of recurrent polyploidization in the Carassius auratus species complex (Cypriniformes, Cyprinidae). *Heredity*, 112(4):415–427.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1):18.

Lynch, M. and Conery, J. S. (2000). The Evolutionary Fate and Consequences of Duplicate Genes. *Science*, 290(5494):1151–1155.

Lynch, M. and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154(1):459–473.

Lyons, E. and Freeling, M. (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant Journal*, 53(4):661–673.

Lyons, E., Freeling, M., Kustu, S., and Inwood, W. (2011). Using genomic sequencing for classical genetics in E. coli K12. *PLoS ONE*, 6(2):e16717.

Lyons, E., Pedersen, B., Kane, J., Alam, M., Ming, R., Tang, H., et al. (2008). Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant physiology*, 148(4):1772–1781.

Ma, J., Li, J., Zhao, J., Zhou, H., Ren, F., Wang, L., et al. (2014). Inactivation of a Gene Encoding Carotenoid Cleavage Dioxygenase (CCD4) Leads to Carotenoid-Based Yellow Coloration of Fruit Flesh and Leaf Midvein in Peach. *Plant Molecular Biology Reporter*, 32(1):246–257.

Madlung, A. (2013). Polyploidy and its effect on evolutionary success: old questions revisited with new tools. *Heredity*, 110(2):99–104.

Madlung, A. and Wendel, J. F. (2013). Genetic and epigenetic aspects of polyploid evolution in plants. *Cytogenetic and Genome Research*, 140(2-4):270–285.

Majoros, W. H., Pertea, M., and Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics (Oxford, England)*, 20(16):2878–9.

Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., and Clavijo, B. J. (2017). KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics (Oxford, England)*, 33(4):574–576.

Marçais, G. and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics (Oxford, England)*, 27(6):764–70.

Marcet-Houben, M. and Gabaldón, T. (2015). Beyond the whole-genome duplication: Phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS Biology*, 13(8):e1002220.

Marchese, A., Marra, F. P., Caruso, T., Mhelembe, K., Costa, F., Fretto, S., and Sargent, D. J. (2016). The first high-density sequence characterized SNP-based linkage map of olive (Olea europaea L. subsp. europaea) developed using genotyping by sequencing. *AJCS*, 10(6):857–863.

Marco-Sola, S., Sammeth, M., Guigó, R., and Ribeca, P. (2012). The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature Methods*, 9(12):1185–1188.

Marcussen, T., Sandve, S. R., Heier, L., Spannagl, M., Pfeifer, M., Jakobsen, K. S., et al. (2014). Ancient hybridizations among the ancestral genomes of bread wheat. *Science*, 345(6194):1250092–1250092.

Margarido, G. R. and Heckerman, D. (2015). ConPADE: genome assembly ploidy estimation from next-generation sequencing data. *PLoS Computational Biology*, 11(4):e1004229.

Margaritis, E. and Jones, M. (2008). Crop processing of Olea europaea L.: An experimental approach for the interpretation of archaeobotanical olive remains. *Vegetation History and Archaeobotany*, 17(4):381–392.

Mariotti, R., Cultrera, N. G. M., Díez, C. M., Baldoni, L., and Rubini, A. (2010). Identification of new polymorphic regions and differentiation of cultivated olives (Olea europaea L.) through plastome sequence comparison. *BMC plant biology*, 10(1):211.

Martelli, G. P., Boscia, D., Porcelli, F., and Saponari, M. (2016). The olive quick decline syndrome in south-east Italy: a threatening phytosanitary emergency.

Martínez-García, P. J., Parfitt, D. E., Bostock, R. M., Fresnedo-Ramírez, J., Vazquez-Lobo, A., Ogundiwin, E. A., et al. (2013). Application of genomic and quantitative genetic tools to identify candidate resistance genes for brown rot resistance in peach. *PLoS ONE*, 8(11):e78634.

Masuelli, R. W., Camadro, E. L., Erazzú, L. E., Bedogni, M. C., and Marfil, C. F. (2009). Homoploid hybridization in the origin and evolution of wild diploid potato species.

Matsuoka, Y. (2011). Evolution of polyploid triticum wheats under cultivation: The role of domestication, natural hybridization and allopolyploid speciation in their diversification.

Mayrose, I., Zhan, S. H., Rothfels, C. J., Arrigo, N., Barker, M. S., Rieseberg, L. H., and Otto, S. P. (2015). Methods for studying polyploid diversification and the dead end hypothesis: A reply to Soltis et al. (2014).

Mayrose, I., Zhan, S. H., Rothfels, C. J., Magnuson-Ford, K., Barker, M. S., Rieseberg, L. H., et al. (2009). Recently Formed Polyploid Plants Diversify at Lower Rates. *American Journal of Botany*, 333(1):1257–1257.

McClintock, B. (1984). The significance of responses of the genome to challenge. *Science*, 226(4676):792–801.

McCouch, S. R. (2001). Genomics and synteny. *Plant physiology*, 125(1):152–155.

McKain, M. R., Tang, H., McNeal, J. R., Ayyampalayam, S., Davis, J. I., DePamphilis, C. W., et al. (2016). A Phylogenomic Assessment of Ancient Polyploidy and Genome Evolution across the Poales. *Genome biology and evolution*, 8(4):1150–1164.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9):1297–303.

Meadows, J. (2005). The Younger Dryas episode and the radiocarbon chronologies of the Lake Huleh and Ghab Valley pollen diagrams, Israel and Syria. *The Holocene*, 15(4):631–636.

Médail, F., Quézel, P., Besnard, G., and Khadari, B. (2001). Systematics , ecology and phylogeographic significance of Olea europaea L . ssp . maroccana ( Greuter & Burdet ) P . Vargas et al ., a relictual olive tree in south-west Morocco. *Botanical Journal of the Linnean Society*, 137(3):249–266.

Mendell, J. E., Clements, K. D., Choat, J. H., and Angert, E. R. (2008). Extreme polyploidy in a large bacterium. *Proceedings of the National Academy of Sciences*, 105(18):6730–6734.

Messing, J. (2009). The polyploid origin of maize. In *Handbook of Maize: Genetics and Genomics*, pages 221–238. Springer New York, New York, NY.

Meudt, H. M., Rojas-Andrés, B. M., Prebble, J. M., Low, E., Garnock-Jones, P. J., and Albach, D. C. (2015). Is genome downsizing associated with diversification in polyploid lineages of Veronica? *Botanical Journal of the Linnean Society*, 178(2):243–266.

Meyer, S., Pospisil, H., and Scholten, S. (2007). Heterosis associated gene expression in maize embryos 6 days after fertilization exhibits additive, dominant and overdominant pattern. *Plant Molecular Biology*, 63(3):381–391.

Minelli, S., Maggini, F., Gelati, M. T., Angiolillo, A., and Cionini, P. G. (2000). The chromosome complement of Olea europaea L.: characterization by differential staining of the chromatin and in-situ hybridization of highly repeated DNA sequences. *Chromosome Research*, 8(7):615–619.

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J. H., et al. (2008). The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature*, 452(7190):991–6.

Mitsui, Y., Shimomura, M., Komatsu, K., Namiki, N., Shibata-Hatta, M., Imai, M., et al. (2015). The radish genome and comprehensive gene expression profile of tuberous root formation and development. *Scientific Reports*, 5(April):10835.

Mkize, N., Hoelmer, K. A., and Villet, M. H. (2008). A survey of fruit-feeding insects and their parasitoids occurring on wild olives, *Olea europaea* ssp. *cuspidata*, in the Eastern Cape of South Africa. *Biocontrol Science and Technology*, 18(10):991–1004.

Monson, R. K. (2003). Gene duplication, neofunctionalization, and the evolution of C 4 photosynthesis. *International Journal of Plant Sciences Int. J. Plant Sci*, 1643(164):43–54.

Montero-Pau, J., Blanca, J., Bombarely, A., Ziarsolo, P., Esteras, C., Martí-Gómez, C., et al. (2017). De-novo assembly of zucchini genome reveals a whole genome duplication associated with the origin of the Cucurbita genus. *bioRxiv*.

Mookerjee, S., Guerin, J., Collins, G., Ford, C., and Sedgley, M. (2005). Paternity analysis using microsatellite markers to identify pollen donors in an olive grove. *Theoretical and Applied Genetics*, 111(6):1174–1182.

Moral, J. and Trapero, A. (2009). Assessing the Susceptibility of Olive Cultivars to Anthracnose Caused by Colletotrichum acutatum. *Plant Disease*, 93(10):1028–1036.

Morelló, J.-R., Romero, M., and Motilva, M.-J. (2004). Effect of the Maturation Process of the Olive Fruit on the Phenolic Fraction of Drupes and Oils from Arbequina, Farga, and Morrut Cultivars. *J. Agric. Food Chem.*, 52(19):6002–6009.

Mou, S., Liu, Z., Guan, D., Qiu, A., Lai, Y., and He, S. (2013). Functional Analysis and Expressional Characterization of Rice Ankyrin Repeat-Containing Protein, OsPIANK1, in Basal Defense against Magnaporthe oryzae Attack. *PLoS ONE*, 8(3):e59699.

Mousavi, S., Mariotti, R., Bagnoli, F., Costantini, L., Cultrera, N., Arzani, K., et al. (2017). The eastern part of the Fertile Crescent concealed an unexpected route of olive (Olea europaea L.) differentiation. *Annals of Botany*, 119(8):1305–1318.

Mower, J. P., Case, A. L., Floro, E. R., and Willis, J. H. (2012). Evidence against equimolarity of large repeat arrangements and a predominant master circle structure of the mitochondrial genome from a monkeyflower (Mimulus guttatus) lineage with cryptic CMS. *Genome Biology and Evolution*, 4(5):670–686.

Muñoz-Mérida, A., González-Plaza, J. J., Cañada, A., Blanco, A. M., Del Carmen García-López, M., Rodríguez, J. M., et al. (2013). De novo assembly and functional annotation of the olive (Olea europaea) transcriptome. *DNA Research*, 20(1):93–108.

Müntzing, A. (1930). Outlines to a genetic monograph of the genus Galeopsis: with special reference to the nature and inheritance of partial sterility. *Hereditas*, 13(2-3):185–341.

Muse, S. V. (1996). Estimating synonymous and nonsynonymous substitution rates. *Mol Biol Evol*, 13(1):105–114.

Myburg, A. A., Grattapaglia, D., Tuskan, G. A., Hellsten, U., Hayes, R. D., Grimwood, J., et al. (2014). The genome of Eucalyptus grandis. *Nature*.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., et al. (2015). Rfam 12.0: updates to the RNA families database. *Nucleic acids research*, 43(Database issue):D130–7.

Nawrocki, E. P. and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics (Oxford, England)*, 29(22):2933–5.

Nei, M. (1973). Analysis of Gene Diversity in Subdivided Populations. *Proceedings of the National Academy of Sciences of the United States of America*, 70(12):3321–3.

Nieto Feliner, G., Álvarez, I., Fuertes-Aguilar, J., Heuertz, M., Marques, I., Moharrek, F., et al. (2017). Is homoploid hybrid speciation that rare? An empiricist's view. *Heredity*, 118(6):513–516.

Ohri, D., Bhargava, A., and Chatterjee, A. (2004). Nuclear DNA amounts in 112 species of tropical hardwoods – new estimates. *Plant biology (Stuttgart, Germany)*, 6(5):555–61.

Olguin, E. R. S., Arrieta-Espinoza, G., Lobo, J. A., and Espinoza-Esquivel, A. M. (2009). Assessment of gene flow from a herbicide-resistant indica rice (Oryza sativa L.) to the Costa Rican weedy rice (Oryza sativa) in Tropical America: Factors affecting hybridization rates and characterization of F1 hybrids. *Transgenic Research*, 18(4):633–647.

Olsson, S., Seoane-Zonjic, P., Bautista, R., Claros, M. G., González-Martínez, S. C., Scotti, I., et al. (2017). Development of genomic tools in a widespread tropical tree, Symphonia globulifera L.f.: a new low-coverage draft genome, SNP and SSR markers. *Molecular Ecology Resources*, 17(4):614–630.

Omrani-Sabbaghi, A., Shahriari, M., Falahati-Anbaran, M., Mohammadi, S. A., Nankali, A., Mardi, M., and Ghareyazie, B. (2007). Microsatellite markers based assessment of genetic diversity in Iranian olive (Olea europaea L.) collections. *Scientia Horticulturae*, 112(4):439–447.

Ozgenturk, N. O., Oru, F., Sezerman, U., Kuçukural, A., Korkut, S. V., Toksoz, F., and Un, C. (2010). Generation and analysis of expressed sequence tags from Olea europaea L. *Comparative and Functional Genomics*, 2010:1–9.

Pafundo, S., Agrimonti, C., and Marmiroli, N. (2005). Traceability of plant contribution in olive oil by amplified fragment length polymorphisms. *Journal of Agricultural and Food Chemistry*, 53(18):6995–7002.

Parisod, C., Holderegger, R., and Brochmann, C. (2010). Evolutionary consequences of autopolyploidy.

Parra, G., Blanco, E., and Guigó, R. (2000). GeneID in Drosophila. *Genome research*, 10(4):511–5.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics (Oxford, England)*, 23(9):1061–7.

Pasqualone, A., Caponio, F., and Blanco, A. (2001). Inter-simple sequence repeat DNA markers for identification of drupes from different Olea europaea L. cultivars. *European Food Research and Technology*, 213(3):240–243.

Paterson, A., Chapman, B., Kissinger, J., Bowers, J., Feltus, F., and Estill, J. (2006). Many gene and domain families have convergent fates following independent whole-genome duplication events in Arabidopsis, Oryza, Saccharomyces and Tetraodon. *Trends in Genetics*, 22(11):597–602.

Paterson, A. H., Wendel, J. F., Gundlach, H., Guo, H., Jenkins, J., Jin, D., et al. (2012). Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. *Nature*, 492(7429):423–427.

Pecoraro, V., Zerulla, K., Lange, C., and Soppa, J. (2011). Quantification of ploidy in proteobacteria revealed the existence of monoploid, (mero-)oligoploid and polyploid species. *PLoS ONE*, 6(1):e16392.

Pellicer, J., Fay, M. F., and Leitch, I. J. (2010). The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society*, 164(1):10–15.

Pelser, P. B., Kennedy, A. H., Tepe, E. J., Shidler, J. B., Nordenstam, B., Kadereit, J. W., and Watson, L. E. (2010). Patterns and causes of incongruence between plastid and nuclear Senecioneae (Asteraceae) phylogenies. *American journal of botany*, 97(5):856–73.

Petit, R. J., Duminil, J., Fineschi, S., Hampe, A., Salvini, D., and Vendramin, G. G. (2005). Comparative organization of chloroplast, mitochondrial and nuclear diversity in plant populations.

Pollard, K. S., Dudoit, S., and Van Der Laan, M. J. (2007). Multiple Testing Procedures: the multtest Package and Applications to Genomics. *Methodology*, 2005(Chapter 15):1–106.

Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, A., and Lareu, M. V. (2013). An overview of STRUCTURE: Applications, parameter settings, and supporting software. *Frontiers in Genetics*, 4(MAY):98.

Powell, A. L. T., Nguyen, C. V., Hill, T., Cheng, K. L., Figueroa-Balderas, R., Aktas, H., et al. (2012). Uniform ripening Encodes a Golden 2-like Transcription Factor Regulating Tomato Fruit Chloroplast Development. *Science*, 336(6089):1711–1715.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–9.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). No Title. *Genetics*, 155(2):945–59.

Proost, S., Bel, M. V., Vaneechoutte, D., Van De Peer, Y., Inzé, D., Mueller-Roeber, B., and Vandepoele, K. (2015). PLAZA 3.0: An access point for plant comparative genomics. *Nucleic Acids Research*, 43(D1):D974–D981.

Purugganan, M. D. and Fuller, D. Q. (2009). The nature of selection during plant domestication. *Nature*, 457(7231):843–8.

Qi, J., Liu, X., Shen, D., Miao, H., Xie, B., Li, X., et al. (2013). A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nature Genetics*, 45(12):1510–1515.

Qin, G., Xu, C., Ming, R., Tang, H., Guyot, R., Kramer, E. M., et al. (2017). The pomegranate (Punica granatum L.) genome and the genomics of punicalagin biosynthesis.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6):841–2.

Raieta, K., Muccillo, L., and Colantuoni, V. (2015). A novel reliable method of DNA extraction from olive oil suitable for molecular traceability. *Food Chemistry*, 172:596–602.

Rallo, P., Dorado, G., and Martin, A. (2000). Development of simple sequence repeats (SSRs) in olive tree (Olea europaea L.). *TAG Theoretical and Applied Genetics*, 101(5-6):984–989.

Ramamurthy, R. K. and Waters, B. M. (2017). Mapping and Characterization of the fefe Gene That Controls Iron Uptake in Melon (Cucumis melo L.). *Frontiers in Plant Science*, 8:1003.

Ramírez-Madera, A. O., Miller, N. D., Spalding, E. P., Weng, Y., and Havey, M. J. (2017). Spontaneous polyploidization in cucumber. *Theoretical and Applied Genetics*, 130(7):1481–1490.

Ramsey, J. and Schemske, D. W. (1998). Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annual Review of Ecology and Systematics*, 29(1):467–501.

Reale, S., Doveri, S., Díaz, A., Angiolillo, A., Lucentini, L., Pilla, F., et al. (2006). SNP-based markers for discriminating olive (Olea europaea L.) cultivars. *Genome*, 49(9):1193–1205.

Refulio-Rodriguez, N. F. and Olmstead, R. G. (2014). Phylogeny of Lamiidae. *American Journal of Botany*, 101(2):287–299.

Renaut, S., Rowe, H. C., Ungerer, M. C., and Rieseberg, L. H. (2014). Genomics of homoploid hybrid speciation: diversity and transcriptional activity of long terminal repeat retrotransposons in hybrid sunflowers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1648):20130345–20130345.

Renner, S. and Zhang, L.-B. (2004). Biogeography of the Pistia Clade (Araceae): Based on Chloroplast and Mitochondrial DNA Sequences and Bayesian Divergence Time Inference. *Systematic Biology*, 53(3):422–432.

Renny-Byfield, S. and Wendel, J. F. (2014). Doubling down on genomes: Polyploidy and crop plants. *American Journal of Botany*, 101(10):1711–1725.

Rensing, S. A., Sheerin, D. J., and Hiltbrunner, A. (2016). Phytochromes: More Than Meets the Eye.

Rhizopoulou, S. (2007). Olea europaea L. A Botanical Contribution to Culture. *American-Eurasian Journal of Agricultural & Environmental Science*, 2(4):382–387.

Rice, A., Glick, L., Abadi, S., Einhorn, M., Kopelman, N. M., Salman-Minkov, A., et al. (2015). The Chromosome Counts Database (CCDB) - a community resource of plant chromosome numbers.

Rieseberg, L. and Wendell, J. (1993). Gene flow and its consequences in plants. *Hybrid zones and the evolutionary process*, pages 70–109.

Rieseberg, L. H. (1997). Hybrid origins of plant species. *Annual Review of Ecology and Systematics*, 28(1):359–389.

Rieseberg, L. H., Kim, S. C., Randell, R. A., Whitney, K. D., Gross, B. L., Lexer, C., and Clay, K. (2007). Hybridization and the colonization of novel habitats by annual sunflowers. *Genetica*, 129(2):149–165.

Rieseberg, L. H., Raymond, O., Rosenthal, D. M., Lai, Z., Livingstone, K., Nakazato, T., et al. (2003). Major ecological transitions in wild sunflowers facilitated by hybridization. *Science (New York, N.Y.)*, 301(5637):1211–6.

Rieseberg, L. H., Whitton, J., and Gardner, K. (1999). Hybrid Zones and the Genetic Architecture of a Barrier to Gene Flow Between Two Sun ower Species. *Biotechnology*, 152(2):713–727.

Rieseberg, L. H., Whitton, J., and Linder, C. R. (1996). Molecular marker incongruence in plant hybrid zones and phylogenetic trees. *Acta Botanica Neerlandica*, 45(3):243–262.

Rieseberg, L. H. and Willis, J. H. (2007a). Plant speciation. *Science (New York, N.Y.)*, 317(5840):910–4.

Rieseberg, L. H. and Willis, J. H. (2007b). Plant Speciation. *Science*, 317(5840):910–914.

Rius, M. and Darling, J. A. (2014). How important is intraspecific genetic admixture to the success of colonising populations? *Trends in ecology & evolution*, 29(4):233–42.

Rodríguez-Santiago, B., Malats, N., Rothman, N., Armengol, L., Garcia-Closas, M., Kogevinas, M., et al. (2010). Mosaic uniparental disomies and aneuploidies as large structural variants of the human genome. *American Journal of Human Genetics*, 87(1):129–138.

Rosenberg, N. A. (2004). DISTRUCT: A program for the graphical display of population structure. *Molecular Ecology Notes*, 4(1):137–138.

Roullier, C., Duputié, A., Wennekes, P., Benoit, L., Fernández Bringas, V. M., Rossel, G., et al. (2013). Disentangling the Origins of Cultivated Sweet Potato (Ipomoea batatas (L.) Lam.). *PLoS ONE*, 8(5):e62707.

Rubio de Casas, R., Besnard, G., Schönswetter, P., Balaguer, L., and Vargas, P. (2006). Extensive gene flow blurs phylogeographic but not phylogenetic signal in Olea europaea L. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, 113(4):575–83.

Ruduś, I., Sasiak, M., and Kepczyński, J. (2013). Regulation of ethylene biosynthesis at the level of 1-aminocyclopropane-1-carboxylate oxidase (ACO) gene.

Rugini, E., Cristofori, V., and Silvestri, C. (2016). Genetic improvement of olive ( Olea europaea L.) by conventional and in vitro biotechnology methods. *Biotechnology Advances*, 34(5):687–696.

Sage, R. F., Sage, T. L., and Kocacinar, F. (2012). Photorespiration and the Evolution of C 4 Photosynthesis. *Annual Review of Plant Biology*, 63(1):19–47.

Şakİroğlu, M. and Brummer, E. C. (2011). Clarifying the ploidy of some accessions in the USDA alfalfa germplasm collection. *Turk J Bot © TÜBTAK*, 35(5):509–519.

Salman-Minkov, A., Sabath, N., and Mayrose, I. (2016). Whole-genome duplication as a key factor in crop domestication. *Nature Plants*, 2(8):16115.

Saminathan, T., Nimmakayala, P., Manohar, S., Malkaram, S., Almeida, A., Cantrell, R., et al. (2015). Differential gene expression and alternative splicing between diploid and tetraploid watermelon. *Journal of Experimental Botany*, 66(5):1369–1385.

Sanderson, M. J. (2003). r8s: Inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, 19(2):301–302.

Saponari, M., Boscia, D., Nigro, F., and Martelli, G. P. (2013). Identification of dna sequences related to Xylella fastidiosa in oleander, almond and olive trees exhibiting leaf scorch symptoms in Apulia (Southern Italy).

Sarah, G., Homa, F., Pointet, S., Contreras, S., Sabot, F., Nabholz, B., et al. (2017). A large set of 26 new reference transcriptomes dedicated to comparative population genomics in crops and wild relatives. *Molecular Ecology Resources*, 17(3):565–580.

Sattler, M. C., Carvalho, C. R., and Clarindo, W. R. (2016). The polyploidy and its key role in plant breeding. *Planta*, 243(2):281–96.

Sauvage, C., Rau, A., Aichholz, C., Chadoeuf, J., Sarah, G., Ruiz, M., et al. (2017). Domestication rewired gene expression and nucleotide diversity patterns in tomato. *The Plant journal : for cell and molecular biology*.

Schäferhoff, B., Fleischmann, A., Fischer, E., Albach, D. C., Borsch, T., Heubl, G., and Müller, K. F. (2010). Towards resolving Lamiales relationships: insights from rapidly evolving chloroplast sequences. *BMC Evolutionary Biology*, 10(1):352.

Schattner, P., Brooks, A. N., and Lowe, T. M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic acids research*, 33(Web Server issue):W686–9.

Schilirò, E., Ferrara, M., Nigro, F., and Mercado-Blanco, J. (2012). Genetic responses induced in olive roots upon colonization by the biocontrol endophytic bacterium Pseudomonas fluorescens PICF7. *PloS one*, 7(11):e48646.

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature*, 463(7278):178–183.

Schwartze, V. U., Winter, S., Shelest, E., Marcet-Houben, M., Horn, F., Wehner, S., et al. (2014). Gene Expansion Shapes Genome Architecture in the Human Pathogen Lichtheimia corymbifera: An Evolutionary Genomics Analysis in the Ancient Terrestrial Mucorales (Mucoromycotina). *PLoS Genetics*, 10(8):e1004496.

Sebastiani, L. and Busconi, M. (2017). Recent developments in olive (Olea europaea L.) genetics and genomics: applications in taxonomy, varietal identification, traceability and breeding. *Plant cell reports*, 36(9):1345–1360.

Sehrish, T., Symonds, V. V., Soltis, D. E., Soltis, P. S., and Tate, J. A. (2014). Gene silencing via DNA methylation in naturally occurring Tragopogon miscellus (Asteraceae) allopolyploids. *BMC Genomics*, 15(701):1–7.

Sémon, M. and Wolfe, K. H. (2007). Consequences of genome duplication.

Session, A. M., Uno, Y., Kwon, T., Chapman, J. A., Toyoda, A., Takahashi, S., et al. (2016). Genome evolution in the allotetraploid frog Xenopus laevis. *Nature*, 538(7625):336–343.

Sheidai, M., Noormohammadi, Z., Dehghani, A., Parvini, F., Hoshiar-Parsian, H., and Hosseini-Mazinani, M. (2010). Intra-specific morphological and molecular diversity in brown olive (Olea cuspidata) of Iran. *ScienceAsia*, 36(3):187–193.

Shoemaker, R. C., Polzin, K., Labate, J., Specht, J., Brummer, E. C., Olson, T., et al. (1996). Genome duplication in soybean (Glycine subgenus soja). *Genetics*, 144(1):329–338.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics (Oxford, England)*, 31(19):3210–2.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome research*, 19(6):1117–23.

Slater, G. S. C. and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics*, 6:31.

Small, R. L., Cronn, R. C., and Wendel, J. F. (2004). Use of nuclear genes for phylogeny reconstruction in plants.

Smit, A., Hubley, R., and Green, P. (2015). RepeatMasker Open-4.0.

Smith, B. (2011). A Cultural Niche Construction Theory of Initial Domestication. *Biological Theory*, 6(3):260–271.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.

Solís-Lemus, C., Ané, C., Michel, B., Slatkin, M., Rokas, A., and Warnow, T. (2016). Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting. *PLOS Genetics*, 12(3):e1005896.

Sollars, E. S. A., Harper, A. L., Kelly, L. J., Sambles, C. M., Ramirez-Gonzalez, R. H., Swarbreck, D., et al. (2016). Genome sequence and genetic diversity of European ash trees. *Nature*, 541(7636):212–216.

Soltis, D. E., Soltis, P. S., Pires, J. C., Kovarik, A., Tate, J. A., and Mavrodiev, E. (2004). Recent and recurrent polyploidy in Tragopogon (Asteraceae): cytogenetic, genomic and genetic comparisons. *Biological Journal of the Linnean Society*, 82(4):485–501.

Soltis, D. E., Visger, C. J., Blaine Marchant, D., and Soltis, P. S. (2016). Polyploidy: Pitfalls and paths to a paradigm. *American Journal of Botany*, 103(7):1146–1166.

Soltis, P. S. (2005). Ancient and recent polyploidy in angiosperms. *New Phytologist*, 166(1):5–8.

Soltis, P. S. and Soltis, D. E. (2009). The Role of Hybridization in Plant Speciation. *Annual Review of Plant Biology*, 60(1):561–588.

Soltis, P. S. and Soltis, D. E. (2016). Ancient WGD events as drivers of key innovations in angiosperms.

Soppa, J. (2011). Ploidy and gene conversion in Archaea. *Biochemical Society Transactions*, 39(1):150–154.

Soppa, J. (2013). Evolutionary advantages of polyploidy in halophilic archaea. *Biochemical Society transactions*, 41(1):339–43.

Soppa, J. (2017). Polyploidy and community structure. *Nature Microbiology*, 2(2):16261.

Spoelhof, J. P., Soltis, P. S., and Soltis, D. E. (2017). Pure polyploidy: Closing the gaps in autopolyploid research.

Spooner, D. M., Rodríguez, F., Polgár, Z., Ballard, H. E., and Jansky, S. H. (2008). Genomic origins of potato polyploids: GBSSI gene sequencing data. *Crop Science*, 48(SUPPL. 1):S–27.

Stace, C. A. (2010). Cytology and Cytogenetics as a Fundamental Taxonomic Resource for the 20th and 21st Centuries Author ( s ): Clive A . Stace Published by : International Association for Plant Taxonomy ( IAPT ) Stable URL : http://www.jstor.org/stable/1224344. *Cytogenetics*, 49(3):451–477.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research*, 34(Web Server issue):W435–9.

Stebbins, G. L. (1971). *Chromosomal evolution in higher plants.* Edward Arnold Ltd., London.

Stebbins, G. L. and Jr. (1938). Cytological Characteristics Associated with the Different Growth Habits in the Dicotyledons. *American Journal of Botany*, 25(3).

Sun, X., Jia, Q., Guo, Y., Zheng, X., and Liang, K. (2015). Whole-genome analysis revealed the positively selected genes during the differentiation of indica and temperate japonica rice. *PLoS ONE*, 10(3):e0119239.

Suzuki, N., Rivero, R. M., Shulaev, V., Blumwald, E., and Mittler, R. (2014). Abiotic and biotic stress combinations. *New Phytologist*, 203(1):32–43.

Talent, N. and Dickinson, T. A. (2005). Polyploidy in Crataegus and Mespilus (Rosaceae, Maloideae): evolutionary inferences from flow cytometry of nuclear DNA amounts. *Canadian Journal of Botany*, 83(10):1268–1304.

Tang, H., Peng, J., Wang, P., and Risch, N. J. (2005). Estimation of individual admixture: Analytical and study design considerations. *Genetic Epidemiology*, 28(4):289–301.

Tang, H., Wang, X., Bowers, J. E., Ming, R., Alam, M., and Paterson, A. H. (2008). Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Research*, 18(12):1944–1954.

Taylor, H. (1945). Cyto-Taxonomy and Phylogeny of the Oleaceae. *Brittonia*, 5(4):337.

Taylor, J. S., Van de Peer, Y., Braasch, I., and Meyer, A. (2001). Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 356(1414):1661–79.

Templeton, A. R. (1981). Mechanisms of Speciation - A Population Genetic Approach. *Annual Review of Ecology and Systematics*, 12(1):23–48.

Terral, J.-F., Alonso, N., Capdevila, R. B. i., Chatti, N., Fabre, L., Fiorentino, G., et al. (2004). Historical biogeography of olive domestication (Olea europaea L.) as revealed by geometrical morphometry applied to biological and archaeological material. *Journal of Biogeography*, 31(1):63–77.

Terzopoulos, P. J., Kolano, B., Bebeli, P. J., Kaltsikes, P. J., and Metzidakis, I. (2005). Identification of Olea europaea L. cultivars using inter-simple sequence repeat markers. *Scientia Horticulturae*, 105(1):45–51.

Than, C., Ruths, D., and Nakhleh, L. (2008). PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, 9(1):322.

Thomas, B. C., Pedersen, B., and Freeling, M. (2006). Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Research*, 16(7):934–946.

Todesco, M., Pascual, M. A., Owens, G. L., Ostevik, K. L., Moyers, B. T., Hübner, S., et al. (2016). Hybridization and extinction. *Evolutionary Applications*, 9(7):892–908.

Tomato Genome Consortium, T. T. G. C. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485(7400):635–41.

Town, C. D. (2006). Comparative Genomics of Brassica oleracea and Arabidopsis thaliana Reveal Gene Loss, Fragmentation, and Dispersal after Polyploidy. *The Plant Cell Online*, 18(6):1348–1359.

Trapero, C., Rallo, L., López-Escudero, F. J., Barranco, D., and Díez, C. M. (2015). Variability and selection of verticillium wilt resistant genotypes in cultivated olive and in the Olea genus. *Plant Pathology*, 64(4):1–11.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–5.

Trujillo, I., Ojeda, M. A., Urdiroz, N. M., Potter, D., Barranco, D., Rallo, L., and Diez, C. M. (2014). Identification of the Worldwide Olive Germplasm Bank of Córdoba (Spain) using SSR and morphological markers. *Tree Genetics and Genomes*, 10(1):141–155.

Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray). *Science*, 313(5793):1596–1604.

Twyford, A. D. and Ennos, R. A. (2012). Next-generation hybridization and introgression. *Heredity*, 108(3):179–189.

Udall, J. A. and Wendel, J. F. (2006). Polyploidy and crop improvement.

Van de Paer, C., Hong-Wa, C., Jeziorski, C., and Besnard, G. (2016). Mitogenomics of Hesperelaea, an extinct genus of Oleaceae. *Gene*, 594(2):197–202.

Van de Peer, Y., Fawcett, J. A., Proost, S., Sterck, L., and Vandepoele, K. (2009a). The flowering world: a tale of duplications. *Trends in Plant Science*, 14(12):680–688.

Van de Peer, Y., Maere, S., and Meyer, A. (2009b). The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics*, 10(10):725–732.

Van de Peer, Y., Mizrachi, E., and Marchal, K. (2017). The evolutionary significance of polyploidy. *Nature Reviews Genetics*, 18(7):411–424.

Van Loo, P., Nordgard, S. H., Lingjaerde, O. C., Russnes, H. G., Rye, I. H., Sun, W., et al. (2010). Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences*, 107(39):16910–16915.

Vanneste, K., Baele, G., Maere, S., and Van de Peer, Y. (2014). Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the CretaceousPaleogene boundary. *Genome Research*, 24(8):1334–1347.

Vanneste, K., Van De Peer, Y., and Maere, S. (2013). Inference of genome duplications from age distributions revisited. *Molecular Biology and Evolution*, 30(1):177–190.

Vargas, P. and Kadereit, J. W. (2001). Molecular fingerprinting evidence (ISSR, inter-simple sequence repeats) for a wild status of Olea europaea L. (Oleaceae) in the Eurosiberian North of the Iberian Peninsula. *Flora*, 196(2):142–152.

Vargas, P., McAllister, H. A., Morton, C., Jury, S. L., and Wilkinson, M. J. (1999). Polyploid speciation inHedera (Araliaceae): Phylogenetic and biogeographic insights based on chromosome counts and ITS sequences. *Plant Systematics and Evolution*, 219(3-4):165–179.

Vargas, P., Muñoz Garmendia, F., Hess, J., and Kadereit, J. (2000). Olea europaea subsp. guanchica and subsp. maroccana (Oleaceae), two new names for olive tree relatives. *Anales del Jardín Botánico de Madrid*, 58(2):360–361.

Vargas, P., Valente, L. M., Blanco-Pastor, J. L., Liberal, I., Guzmán, B., Cano, E., Forrest, A., and Fernández-Mazuecos, M. (2014). Testing the biogeographical congruence of palaeofloras using molecular phylogenetics: snapdragons and the Madrean-Tethyan flora. *Journal of Biogeography*, 41(5):932–943.

Vargas, P. and Zardoya, R., editors (2012). *El árbol de la vida: sistemática y evolución de los seres vivos*. Madrid, cuarta impresión edition.

Vargas, P. and Zardoya, R. (2014). *The tree of life : evolution and classification of living organisms*.

Vegas, J., Garcia-Mas, J., and Monforte, A. J. (2013). Interaction between QTLs induces an advance in ethylene biosynthesis during melon fruit ripening. *Theoretical and Applied Genetics*, 126(6):1531–1544.

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., et al. (2010). The genome of the domesticated apple (Malus domestica Borkh.). *Nature Genetics*, 42(10):833–839.

Veltri, D., Wight, M. M., and Crouch, J. A. (2016). SimpleSynteny: a web-based tool for visualization of microsynteny across multiple species. *Nucleic Acids Research*, 44(May):gkw330.

Vendramin, E., Pea, G., Dondini, L., Pacheco, I., Dettori, M. T., Gazza, L., et al. (2014). A unique mutation in a MYB gene cosegregates with the nectarine phenotype in peach. *PLoS ONE*, 9(3):e90574.

Vining, K. J., Johnson, S. R., Ahkami, A., Lange, I., Parrish, A. N., Trapp, S. C., et al. (2017). Draft Genome Sequence of Mentha longifolia and Development of Resources for Mint Cultivar Improvement. *Molecular Plant*, 10(2):323–339.

Vision, T. J., Brown, D. G., and Tanksley, S. D. (2000). The Origins of Genomic Duplications in Arabidopsis. *Science*, 290(5499):2114–2117.

Vlasova, A., Capella-Gutiérrez, S., Rendón-Anaya, M., Hernández-Oñate, M., Minoche, A. E., Erb, I., et al. (2016). Genome and transcriptome analysis of the Mesoamerican common bean and the role of gene duplications in establishing tissue and temporal specialization of genes. *Genome Biology*, 17(1):32.

Vo, K. T. X., Kim, C.-Y., Chandran, A. K. N., Jung, K.-H., An, G., and Jeon, J.-S. (2015). Molecular insights into the function of ankyrin proteins in plants. *Journal of Plant Biology*, 58(5):271–284.

Wallace, I. M., O'Sullivan, O., Higgins, D. G., and Notredame, C. (2006). M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic acids research*, 34(6):1692–9.

Wallander, E. (2008). Systematics of Fraxinus (Oleaceae) and evolution of dioecy. *Plant Systematics and Evolution*, 273(1-2):25–49.

Wallander, E. and Albert, V. A. (2000). Phylogeny and classification of Oleaceae based on rps16 and trnL-F sequence data. *American journal of botany*, 87(12):1827–41.

Walsh, J. B. (1995). How often do duplicated genes evolve new functions? *Genetics*, 139(1):421–428.

Wang, K., Wang, Z., Li, F., Ye, W., Wang, J., Song, G., et al. (2012). The draft genome of a diploid cotton Gossypium raimondii. *Nature Genetics*, 44(10):1098–1103.

Wang, L., ting Ge, T., tao Peng, H., Wang, C., kun Liu, T., lin Hou, X., and Li, Y. (2013). Molecular cloning, expression analysis and localization of Exo70A1 related to self incompatibility in non-heading Chinese cabbage (Brassica campestris ssp. chinensis). *Journal of Integrative Agriculture*, 12(12):2149–2156.

Wang, L., Xia, Q., Zhang, Y., Zhu, X., Zhu, X., Li, D., et al. (2016a). Updated sesame genome assembly and fine mapping of plant height and seed coat color QTLs using a new high-density genetic map. *BMC Genomics*, 17(1):31.

Wang, L., Yu, S., Tong, C., Zhao, Y., Liu, Y., Song, C., et al. (2014). Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis. *Genome Biology*, 15(2):R39.

Wang, L. S., Li, W. L., Qi, X. W., Ma, L., and Wu, W. L. (2017a). Physiological and proteomic response of Limonium bicolor to salinity. *Russian Journal of Plant Physiology*, 64(3):349–360.

Wang, X., Guo, H., Wang, J., Lei, T., Liu, T., Wang, Z., et al. (2016b). Comparative genomic de-convolution of the cotton genome revealed a decaploid ancestor and widespread chromosomal fractionation. *New Phytologist*, 209(3):1252–1263.

Wang, X. and Szmidt, A. (1990). Evolutionary analysis of Pinus densata (Masters), a putative Tertiary hybrid. 2. A study using species-specific chloroplast DNA markers. *Theor Appl Genet*, 80(5):641–647.

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., et al. (2011). The genome of the mesopolyploid crop species Brassica rapa. *Nature Genetics*, 43(10):1035–1039.

Wang, X., Xu, Y., Zhang, S., Cao, L., Huang, Y., Cheng, J., et al. (2017b). Genomic analyses of primitive, wild and cultivated citrus provide insights into asexual reproduction. *Nature Genetics*, 49(5):765–772.

Wang, Z., Ge, Q., Chen, C., Jin, X., Cao, X., and Wang, Z. (2017c). Function Analysis of Caffeoyl-CoA O-Methyltransferase for Biosynthesis of Lignin and Phenolic Acid in Salvia miltiorrhiza. *Applied Biochemistry and Biotechnology*, 181(2):562–572.

Watanabe, K. (2002). Index to chromosome numbers in Asteraceae.

Wehe, A., Bansal, M. S., Burleigh, J. G., and Eulenstein, O. (2008). DupTree: A program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics*, 24(13):1540–1541.

Weiß, C. L., Pais, M., Cano, L. M., and Kamoun, S. (2017). nQuire : A statistical framework for ploidy estimation using next generation sequencing. *bioRxiv*, pages 3–7.

Weiss, E. (2015). Beginnings of Fruit Growing in the Old World two generations later. *Israel Journal of Plant Sciences*, 62(1-2):75–85.

Wendel, J. F. (2000). Plant Molecular Evolution. In *Plant Molecular Evolution*, pages 225–249. Springer Netherlands, Dordrecht.

Wendel, J. F. and Cronn, R. C. (2001). Polyploidy and the evolutionary history of cotton.

Wicke, S., Schneeweiss, G. M., DePamphilis, C. W., Müller, K. F., and Quandt, D. (2011). The evolution of the plastid chromosome in land plants: Gene content, gene order, gene function.

Wickham, H. (2009). *Ggplot2 : elegrant graphics for data analysis*. Springer-Verlag, New York, NY.

Wiesman, Z., Avidan, N., Lavee, S., and Quebedeaux, B. (1998). Molecular characterization of common olive varieties in Israel and the West Bank using randomly amplified polymorphic DNA (RAPD) markers.

Williams, D. A., Muchugu, E., Overholt, W. A., and Cuda, J. P. (2007). Colonization patterns of the invasive Brazilian peppertree, Schinus terebinthifolius, in Florida. *Heredity*, 98(5):284–293.

Willyard, A., Cronn, R., and Liston, A. (2009). Reticulate evolution and incomplete lineage sorting among the ponderosa pines. *Molecular Phylogenetics and Evolution*, 52(2):498–511.

Wolfe, K. H. (2001). Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics*, 2(5):333–341.

Wolfe, K. H. (2015). Origin of the yeast whole-genome duplication. *PLoS Biology*, 13(8):e1002221.

Wood, T. E., Takebayashi, N., Barker, M. S., Mayrose, I., Greenspoon, P. B., and Rieseberg, L. H. (2009). The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences*, 106(33):13875–13879.

Wortley, A. H., Rudall, P. J., Harris, D. J., and Scotland, R. W. (2005). How Much Data are Needed to Resolve a Difficult Phylogeny? Case Study in Lamiales. *Systematic Biology*, 54(5):697–709.

Wu, S.-B., Collins, G., and Sedgley, M. (2004). A molecular linkage map of olive (Olea europaea L) based on RAPD, microsatellite, and SCAR markers. *Genome / National Research Council Canada = Genome / Conseil national de recherches Canada*, 47(1):26–35.

Wu, T. D. and Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics (Oxford, England)*, 21(9):1859–75.

Wyman, S. K., Jansen, R. K., and Boore, J. L. (2004). Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*, 20(17):3252–3255.

Xu, B. and Yang, Z. (2013). PamlX: A graphical user interface for PAML. *Molecular Biology and Evolution*, 30(12):2723–2724.

Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., et al. (2011). Genome sequence and analysis of the tuber crop potato. *Nature*, 475(7355):189–95.

Yakimowski, S. B. and Rieseberg, L. H. (2014). The role of homoploid hybridization in evolution: A century of studies synthesizing genetics and ecology. *American Journal of Botany*, 101(8):1247–1258.

Yan, H., Martin, S. L., Bekele, W. A., Latta, R. G., Diederichsen, A., Peng, Y., and Tinker, N. A. (2016). Genome size variation in the genus Avena. *http://dx.doi.org/10.1139/gen-2015-0132*, 59(3):209–220.

Yang, J., Grünewald, S., Xu, Y., and Wan, X.-F. (2014). Quartet-based methods to reconstruct phylogenetic networks. *BMC Systems Biology*, 8(1):21.

Yang, X., Ye, C.-Y., Cheng, Z.-M., Tschaplinski, T. J., Wullschleger, S. D., Yin, W., et al. (2011). Genomic aspects of research involving polyploid plants. *Plant Cell, Tissue and Organ Culture (PCTOC)*, 104(3):387–397.

Yang, Y., Moore, M. J., Brockington, S. F., Soltis, D. E., Wong, G. K. S., Carpenter, E. J., et al. (2015). Dissecting molecular evolution in the highly diverse plant clade caryophyllales using transcriptome sequencing. *Molecular Biology and Evolution*, 32(8):2001–2014.

Yoruk, B. and Taskin, V. (2014). Genetic diversity and relationships of wild and cultivated olives in Turkey. *Plant Systematics and Evolution*, 300(5):1247–1258.

Yu, J. (2002). A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. indica). *Science*, 296(5565):79–92.

Yu, J., Wang, L., Guo, H., Liao, B., King, G., and Zhang, X. (2017). Genome evolutionary dynamics followed by diversifying selection explains the complexity of the Sesamum indicum genome. *BMC genomics*, 18(1):257.

Zachos, J. (2001). Trends, Rhythms, and Aberrations in Global Climate 65 Ma to Present. *Science*, 292(5517):686–693.

Zachos, J. C., Shackleton, N. J., Revenaugh, J. S., Pälike, H., and Flower, B. P. (2001). Climate response to orbital forcing across the Oligocene-Miocene boundary. *Science (New York, N.Y.)*, 292(5515):274–8.

Zahn, L. M., Leebens-Mack, J. H., Arrington, J. M., Hu, Y., Landherr, L. L., DePamphilis, C. W., et al. (2006). Conservation and divergence in the AGAMOUS subfamily of MADS-box genes: Evidence of independent sub- and neofunctionalization events. *Evolution and Development*, 8(1):30–45.

Zeder, M. A. (2015). Core questions in domestication research. *Proceedings of the National Academy of Sciences*, 112(11):3191–3198.

Zenil-Ferguson, R., Ponciano, J. M., and Burleigh, J. G. (2017). Testing the association of phenotypes with polyploidy: An example using herbaceous and woody eudicots. *Evolution*, 71(5):1138–1148.

Zhan, S. H., Drori, M., Goldberg, E. E., Otto, S. P., and Mayrose, I. (2016). Phylogenetic evidence for cladogenetic polyploidization in land plants. *American journal of botany*, 103(7):1252–8.

Zhang, C., Li, J., Guo, X., Zhu, B., Xiao, W., Wang, P., et al. (2017a). LecRK-VII.1, a Lectin Receptor-Like Kinase, Mediates the Regulation of Salt Stress and Jasmonic Acid Response in Arabidopsis. *Journal of Plant Growth Regulation*, 36(2):385–401.

Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., et al. (2012a). The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*, 490(7418):49–54.

Zhang, G., Tian, Y., Zhang, J., Shu, L., Yang, S., Wang, W., et al. (2015a). Hybrid de novo genome assembly of the Chinese herbal plant danshen (Salvia miltiorrhiza Bunge). *GigaScience*, 4:62.

Zhang, H., Miao, H., Wang, L., Qu, L., Liu, H., Wang, Q., and Yue, M. (2013). Genome sequencing of the important oilseed crop Sesamum indicum L. *Genome biology*, 14(1):401.

Zhang, J., Tian, Y., Yan, L., Zhang, G., Wang, X., Zeng, Y., et al. (2016). Genome of Plant Maca (Lepidium meyenii) Illuminates Genomic Basis for High-Altitude Adaptation in the Central Andes. *Molecular Plant*, 9(7):1066–1077.

Zhang, N., Erickson, D. L., Ramachandran, P., Ottesen, A. R., Timme, R. E., Funk, V. A., et al. (2017b). An analysis of Echinacea chloroplast genomes: Implications for future botanical identification. *Scientific Reports*, 7(1):216.

Zhang, Q., Li, J., Zhao, Y., Korban, S. S., and Han, Y. (2012b). Evaluation of Genetic Diversity in Chinese Wild Apple Species Along with Apple Cultivars Using SSR Markers. *Plant Molecular Biology Reporter*, 30(3):539–546.

Zhang, T., Hu, Y., Jiang, W., Fang, L., Guan, X., Chen, J., et al. (2015b). Sequencing of allotetraploid cotton (Gossypium hirsutum L. acc. TM-1) provides a resource for fiber improvement. *Nature Biotechnology*, 33(5):531–537.

Zhang, Z., Belcram, H., Gornicki, P., Charles, M., Just, J., Huneau, C., et al. (2011). Duplication and partitioning in evolution and function of homoeologous Q loci governing domestication characters in polyploid wheat. *Proceedings of the National Academy of Sciences of the United States of America*, 108(46):18737–42.

Zhou, Y., Massonnet, M., Sanjak, J., Cantu, D., and Gaut, B. S. (2017). The Evolutionary Genomics of Grape (Vitis vinifera ssp. vinifera) Domestication. *bioRxiv*.

Zhu, Z., Xu, F., Zhang, Y., Cheng, Y. T., Wiermer, M., Li, X., and Zhang, Y. (2010). Arabidopsis resistance protein SNC1 activates immune responses through association with a transcriptional corepressor. *Proceedings of the National Academy of Sciences of the United States of America*, 107(31):13960–13965.

Zohary, D., Hopf, M., and Weiss, E. (2012). *Domestication of plants in the Old World: the origin and spread of domesticated plants in Southwest Asia, Europe, and the Mediterranean Basin*. Oxford University Press.

Zohary, D. and Spiegel-Roy, P. (1975). Beginnings of fruit growing in the Old World. *Science, USA*, 187(4174):319–327.

Zohren, J., Wang, N., Kardailsky, I., Borrell, J. S., Joecker, A., Nichols, R. A., and Buggs, R. J. (2016). Unidirectional diploidtetraploid introgression among British birch trees with shifting ranges shown by restriction site-associated markers. *Molecular Ecology*, 25(11):2413–2426.