

RESEARCH ARTICLE

Open Access



Phylogenomics of the *Olea europaea* complex using 15 whole genomes supports recurrent genetic admixture together with differentiation into seven subspecies

Irene Julca^{1,2,3}, Pablo Vargas⁴ and Toni Gabaldón^{1,2,5,6*} 

Abstract

Background The last taxonomic account of *Olea* recognises six subspecies within *Olea europaea* L., including the Mediterranean olive tree (subsp. *europaea*) and five other subspecies (*laperrinei*, *guanchica*, *maroccana*, *cerasiformis*, and *cuspidata*) distributed across the Old World, including Macaronesian islands. The evolutionary history of this monophyletic group (*O. europaea* complex) has revealed a reticulated scenario involving hybridization and polyploidization events, leading to the presence of a polyploid series associated with the subspecies. However, how the polyploids originated, and how the different subspecies contributed to the domestication of the cultivated olive are questions still debated. Tracing the recent evolution and genetic diversification of the species is key for the management and preservation of its genetic resources. To study the recent history of the *O. europaea* complex, we compared newly sequenced and available genomes for 27 individuals representing the six subspecies.

Results Our results show discordance between current subspecies distributions and phylogenomic patterns, which support intricate biogeographic patterns. The subspecies *guanchica*, restricted to the Canary Islands, is closely related to subsp. *europaea*, and shows a high genetic diversity. The subsp. *laperrinei*, restricted now to high mountains of the Sahara desert, and the Canarian subsp. *guanchica* contributed to the formation of the allotetraploid subsp. *cerasiformis* (Madeira islands) and the allohexaploid subsp. *maroccana* (western Sahara region). Our phylogenomic data support the recognition of one more taxon (subsp. *ferruginea*) for the Asian populations, which is clearly segregated from the African subsp. *cuspidata*.

Conclusions In sum, the *O. europaea* complex underwent several processes of hybridization, polyploidy, and geographical isolation resulting in seven independent lineages with certain morphological traits recognised into subspecies.

Keywords Macaronesian islands, Olive (*Olea europaea*), Phylogenomics, Polyploidy, Subspecies

*Correspondence:

Toni Gabaldón
toni.gabaldon@bsc.es

Full list of author information is available at the end of the article



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

The *Olea europaea* complex [1] includes the wild and cultivated Mediterranean olives (*O. europaea* subspecies *europaea*), and five non-Mediterranean subspecies (Fig. 1a): subsp. *laperrinei* distributed in Saharan massifs; subsp. *cuspidata*, with a broad distribution ranging from South Africa to southern Egypt and from Arabia to northern India and southwestern China; subsp.

guanchica endemic to the Canary Islands; subsp. *maroccana* in the Agadir Mountains (western Sahara region); and subsp. *cerasiformis* in the Madeira islands [2, 3]. The subsp. *europaea* is further subdivided into two taxonomic varieties: var. *sylvestris*, also named oleaster or wild olive, which comprises the wild forms of the olive tree, and var. *europaea*, which encompasses around 1,000 cultivated forms [2, 4]. Recent analyses of the first complete

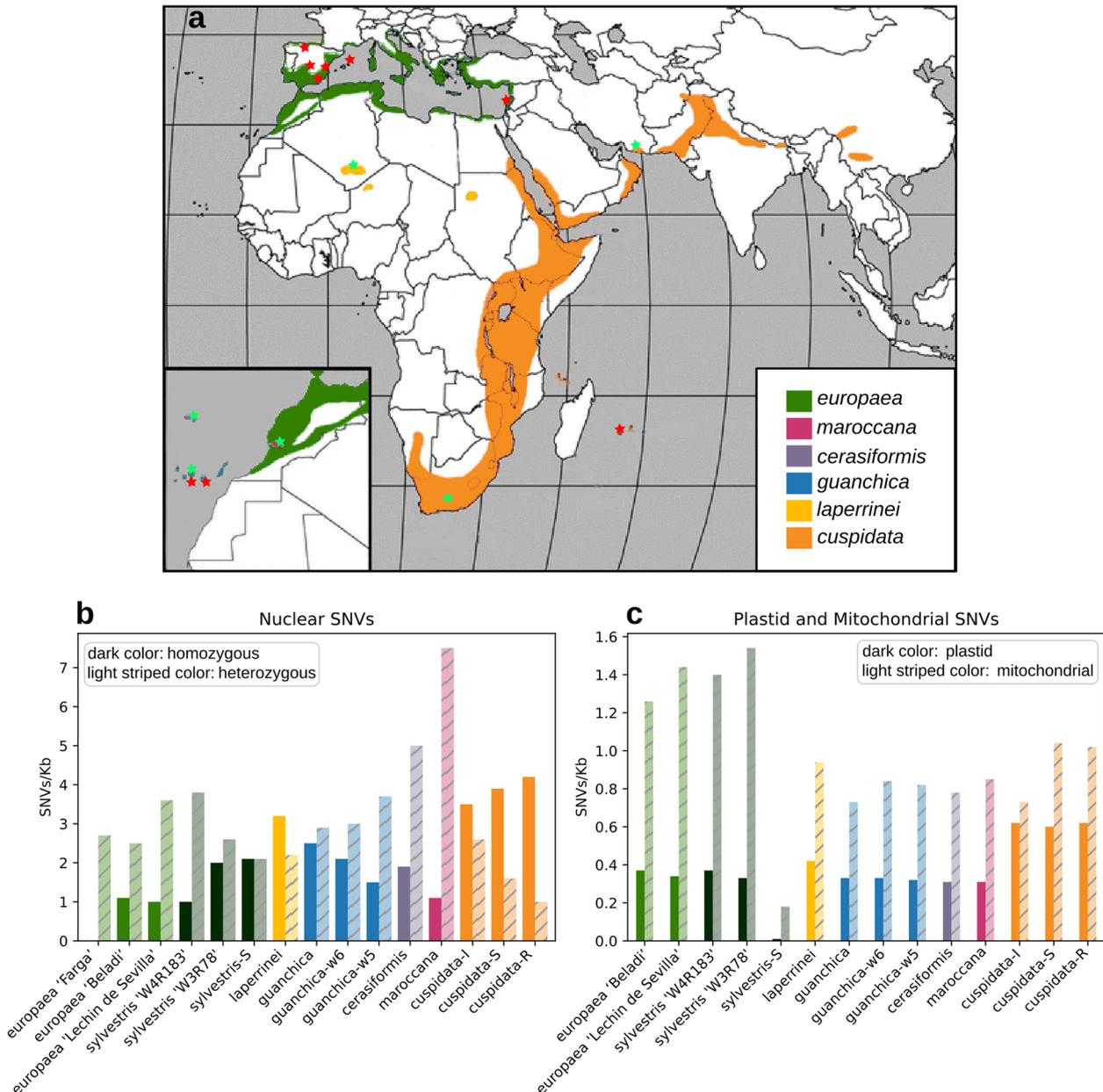


Fig. 1 *Olea europaea* sampling and SNV densities (SNVs/Kb). **a** Geographical distribution of the six *O. europaea* subspecies (modified from [8]). Samples sequenced in this project are marked with green stars and other whole-genome sequences with red stars. Colours show the distribution of each subspecies. **b** SNVs/Kb for homozygous (darker colours) and heterozygous (lighter striped colours) SNVs in the nuclear genome. **c** SNV densities for the plastid (dark colour) and mitochondrial (light striped colour) genomes

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 [5, 6]. In addition to these ancient events, earlier work had described hybridization processes within the *O. europaea* complex [4, 7, 8] and had identified the presence of a polyploid series, including some derived diploid subspecies (*europaea*, *laperrinei*, *cuspidata*, *guanchica*) with $2n=46$, the tetraploid subsp. *cerasiformis*, and the hexaploid subsp. *maroccana* [1, 9]. However, these studies are limited because of the use of a low sample size in terms of the number of genetic markers. Therefore, the question remains as to how the polyploids originated, and how the different subspecies contributed to the domestication of the cultivated olive. In addition, a recent genomic analysis of cultivars and wild relatives has identified common introgression and hybridization [10], but the extent to which such processes have operated in the earlier diversification of the subspecies remains to be disclosed. Access to genome sequences of additional subspecies of the *O. europaea* complex is necessary to understand the recent evolution of the species and to assess the genomic aftermath of past processes of isolation, hybridization and whole-genome duplication. To infer the recent evolution of the *O. europaea* complex, we whole sequenced one individual of each of the following subspecies: *cerasiformis*, *maroccana*, *guanchica*, *laperrinei*, and two individuals of the subspecies *cuspidata* from two continents (Africa, Asia). We integrated these sequenced individuals with available genome sequences of the subspecies *europaea*, *guanchica*, and *cuspidata*. A comparative analysis of a total of 27 olive genomes is addressed to shed further light on the evolutionary history of the *O. europaea* complex.

Results

Patterns of genetic polymorphism in *O. europaea*

To study patterns of genetic polymorphism within *Olea europaea*, we sequenced six individuals of the five subspecies of the *O. europaea* complex and retrieved available sequences from ENA and NCBI databases: nuclear genome sequences for nine individuals of subsp. *guanchica*, *europaea* and *cuspidata*, and 12 plastid sequences from all six subspecies. These samples represent the current geographic distribution of the species (Fig. 1a, Table 1). Using a read mapping strategy (see [Methods](#)), we called single nucleotide variants (SNVs), which resulted in a total of 27,877,747 nuclear, 3,385 mitochondrial, and 238 plastidial polymorphic positions. The number of identified homozygous nuclear SNVs per individual shows that subsp. *cuspidata* has the largest numbers (~3.9 SNVs/Kb), followed by subsp. *laperrinei* (3.2 SNVs/Kb) (Fig. 1b). The other subspecies have

a similar amount of homozygous SNVs with respect to the reference. Interestingly, the two polyploid subspecies [9] *maroccana* (6x), and *cerasiformis* (4x) showed significantly higher amounts of heterozygous SNVs as compared to the other individuals, suggesting hybridization events (including introgression) have been involved in the origin of their polyploidy. The analysis of plastid SNVs (Fig. 1c, darker colours) resulted in a similar ordering of genetic distances to the reference of the different subspecies (*cuspidata* > *laperrinei* > other). However, for mitochondrial SNVs (Fig. 1c, lighter striped colours) the subspecies *europaea* shows a higher number of SNVs than the other subspecies. The individual with the lowest number of plastid and mitochondrial polymorphic positions is the subsp. *europaea* var. *sylvestris*—S, as already noted in our previous study [10].

To study structural changes within *O. europaea*, we focused on small insertions/deletions (indels). We identified a total of 1,504,957 indel positions that have sizes ranging from 1 to 313 nucleotides. Interestingly, the two polyploid individuals (subsp. *laperrinei* and *maroccana*) have the highest number of total indels and large indels (>100 nucleotides) (Additional file 1: Table S1). When we searched for indels present in genic regions, we found 424,944 positions affecting genes. The number of genes affected by indels are similar for all individuals, with an average of 25,627 genes (Additional file 1: Table S2). GO term functional enrichment analysis shows enrichment for nucleus and cytoplasm cellular component, microtubule-based movement biological process, and different molecular functions, including microtubule motor activity, microtubule binding, phosphatase activity, palmitoyltransferase activity, ATP binding, and threonine-type endopeptidase activity (Additional file 1: Table S2). Interestingly, palmitoyltransferase activity is a key enzyme of sphingolipid metabolism [12], which has been shown to play an important role during olive fruit development and ripening [13, 14].

Patterns of allelic representation in heterozygous positions

To assess the ploidy of each sequenced individual we plotted the relative coverage of alternative alleles in heterozygous sites (see [Methods](#)). In these plots (Fig. 2) polyploid subspecies show allele frequency peaks to a certain extent consistent with their polyploid level: hexaploid (five peaks at 0.17, 0.33, 0.50, 0.67, 0.83) in subsp. *maroccana* and tetraploid (three peaks at 0.25, 0.50, 0.75) in subsp. *cerasiformis* [9]. For the diploid subspecies, we observed a prominent peak at the expected frequency of 0.50. In addition, a second peak, much broader and of lower height, was observed around frequencies between 0.05 to 0.25. As this secondary peak also appears in the reference, it may represent segmental duplications,

Table 1 Twenty-seven genomes of *O. europaea* used in the analysis. Columns show taxon, geographical origin, plastid group [11], ploidy level, genomes analysed, and data source

Taxon (subspecies)	Geographic origin	Plastid group	Ploidy level	Genome analysed	Source
<i>europaea</i> var. <i>europaea</i> cv. 'Farga'	Spain (Boadilla/La Senia)	E3.1	2x	nuclear, plastid, mitochondrial	ENA (PRJEB4992)
<i>europaea</i> var. <i>europaea</i> cv. 'Lechin de Sevilla'	Spain	E2.3	2x	nuclear, plastid, mitochondrial	ENA (PRJEB35540)
<i>europaea</i> var. <i>europaea</i> cv. 'Beladi'	Lebanon	E1.1	2x	nuclear, plastid, mitochondrial	ENA (PRJEB35540)
<i>europaea</i> var. <i>sylvestris</i>	Spain (Pechón)	E3	2x	nuclear, plastid, mitochondrial	ENA (PRJEB35540)
<i>europaea</i> var. <i>sylvestris</i> 'W4R183'	Spain	E1	2x	nuclear, plastid, mitochondrial	NCBI (SRR9860503)
<i>europaea</i> var. <i>sylvestris</i> 'W3R78'	Spain (Menorca)	E2	2x	nuclear, plastid, mitochondrial	NCBI (SRR9860505)
<i>europaea</i> var. <i>sylvestris</i> 'Stavrovouni 11'	Cyprus	E1.4	2x	plastid	NCBI (HF558645)
<i>europaea</i> var. <i>sylvestris</i> 'Haut Atlas 1'	Morocco (High Atlas)	E2	2x	plastid	NCBI (NC_015401)
<i>europaea</i> var. <i>sylvestris</i> 'Gue de Constantine 20'	Algeria (Gue de Constantine)	E3	2x	plastid	NCBI (FN997651)
<i>europaea</i> var. <i>sylvestris</i> 'Oeirás 1'	Portugal	E3.1	2x	plastid	NCBI (MG255763)
<i>europaea</i> var. <i>sylvestris</i> 'Vallee du Fango 5'	France	E2.1	2x	plastid	NCBI (MG255762)
<i>maroccana</i> ^a	Morocco (Agadir)	M	6x	nuclear, plastid, mitochondrial	Olive genome project
<i>maroccana</i> 'Immouzzet S1'	Morocco (High Atlas)	M	6x	plastid	NCBI (NC_015623)
<i>cerasiformis</i> ^a	Portugal (Madeira)	M	4x	nuclear, plastid, mitochondrial	Olive genome project
<i>guanchica</i> ^a	Spain (Tenerife)	M	2x	nuclear, plastid, mitochondrial	Olive genome project
<i>guanchica</i> 'La Gomera 10'	Spain	M-g1	2x	plastid	NCBI (MG255764)
<i>guanchica</i> -w6	Spain (Tenerife)	M	2x	nuclear, plastid, mitochondrial	NCBI (SRR9860513)
<i>guanchica</i> -w5	Spain (Gran Canaria)	M	2x	nuclear, plastid, mitochondrial	NCBI (SRR9860515)
<i>laperrinei</i> ^a	Sahara	E1.1	2x	nuclear, plastid, mitochondrial	Olive genome project
<i>laperrinei</i> 'Adjelella 10'	Algeria	E1-11	2x	plastid	NCBI (MG255765)
<i>cuspidata</i> -R	Reunion island	A	2x	nuclear, plastid, mitochondrial	ENA (PRJEB35540)
<i>cuspidata</i> -S ^a	South Africa	A	2x	nuclear, plastid, mitochondrial	Olive genome project
<i>cuspidata</i> -I ^a	Iran	C1	2x	nuclear, plastid, mitochondrial	Olive genome project
<i>cuspidata</i> 'Almihwit 5.1'	Yemen	C2	2x	plastid	NCBI (FN996943)
<i>cuspidata</i> 'Guangzhou 1'	China	C1	2x	plastid	NCBI (FN996944)
<i>cuspidata</i> 'Maui 1'	USA (Hawaii-Maui)	A	2x	plastid	NCBI (NC_015604)
<i>cuspidata</i> 'Menagesha Forest 14'	Ethiopia	C2	2x	plastid	NCBI (MG255760)

^a Samples sequenced in this project

aneuploidies, or repetitive regions that may be collapsed in the reference assembly. Differences in the height of the secondary peak could represent variations in the amount of such repetitive regions in other individuals. Interestingly, the subsp. *cuspidata* shows different patterns for the three individuals with *cuspidata*-I from Iran being diploid as previously reported [15], and the other two individuals (*cuspidata*-R and *cuspidata*-S) from Reunion Island and South Africa showing a higher secondary peak, which could reflect differences in genome size previously observed in this subspecies [9].

In order to determine whether repetitive regions are responsible for the secondary peak, we selected all the positions with values ranging from 0.05 to 0.25 (P25) and analysed if they fall inside repetitive regions or genes.

Interestingly, an average of 58% of such positions falls in repetitive regions, while 21% falls in genes (Additional file 1: Table S3). The individuals with more P25 positions (>60%) in repetitive regions are *cuspidata*-S (61%), *cuspidata*-R (63%), *sylvestris*-S (61%), and *sylvestris* 'W3R78' (63%), while the polyploid individuals show the lowest values (*maroccana*—44% and *cerasiformis*—46%). For the *cuspidata* subspecies, we observe that the individual that does not have a clear secondary peak, *cuspidata*-I, has the lowest percentage (57%) of P25 positions falling in repetitive regions. Additionally, when we performed a GO term enrichment analysis of the genes that contain the P25 positions, we found that ATP binding is enriched in the three *cuspidata* individuals, *laperrinei*, and *sylvestris* 'W3R78' (Additional file 1: Table S3). Interestingly,

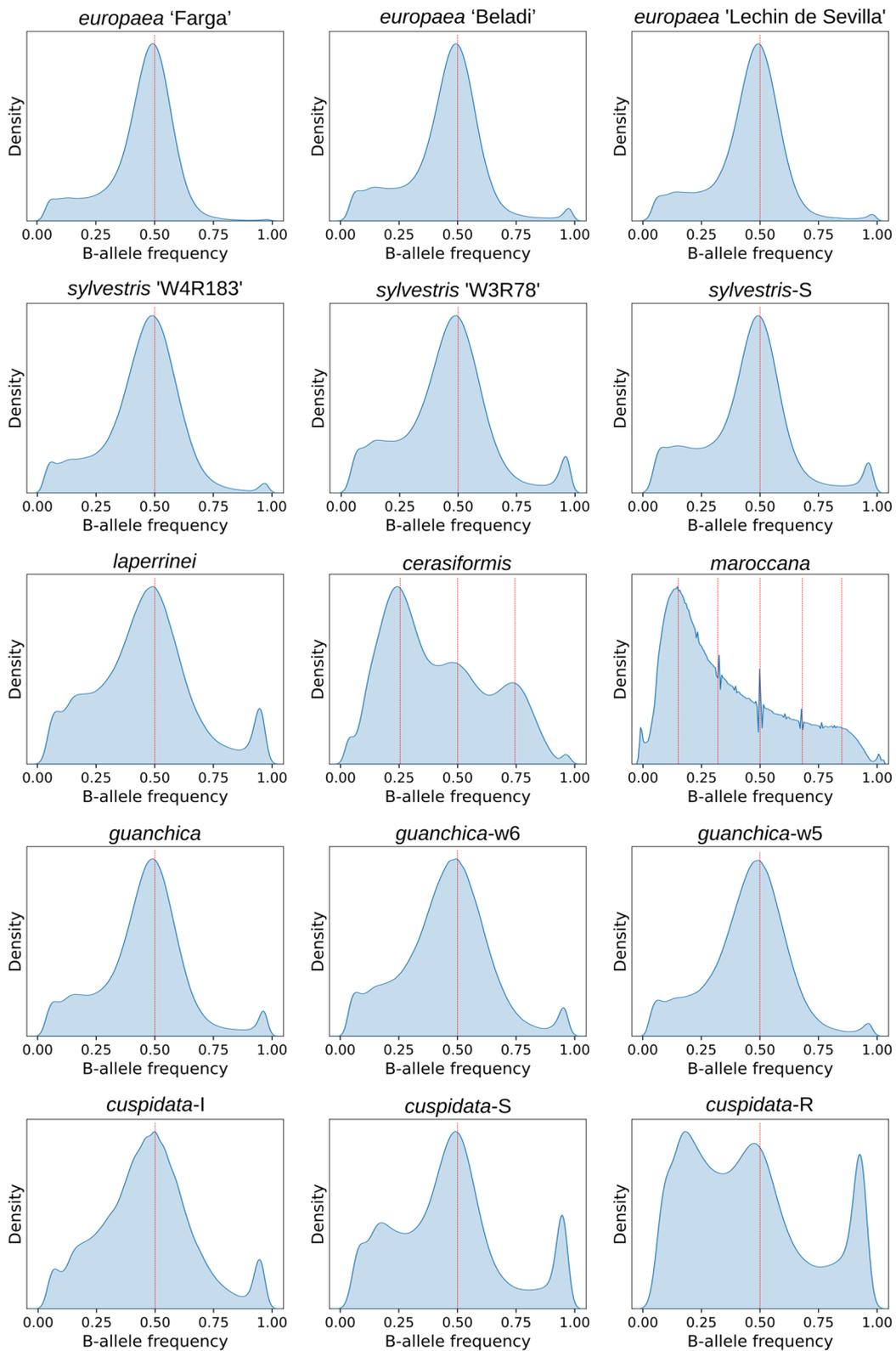


Fig. 2 Density plot for the relative coverage of alternative alleles in heterozygous sites per individual. For all cases, we only plotted data corresponding to the 23 pseudo-chromosomes. Red line indicates the position of the peak consistent with the ploidy

ATP-binding cassette transporters in plants are a very large and diverse protein family and their number can vary even within the same species [16–18]. Altogether, these results show evidence that this secondary peak may be a product of repetitive regions and segmental duplications containing genes.

Genetic diversity in the *O. europaea* complex

Interestingly, small populations (such as subsp. *laperrinei*) and populations in oceanic islands (subsp. *guanchica*) are associated with decreased levels of genetic diversity [19, 20]. To test this, we calculated the genetic diversity of 4 groups of the *O. europaea* complex (Table 2). For the subsp. *europaea*, we analysed the two varieties independently and observed a slightly higher genetic diversity in var. *sylvestris*, which is in agreement with previous studies [10, 11, 21]. When we compare the three subspecies we can observe that subsp. *guanchica* has the highest genetic diversity (Nei, $H = 2.8 \times 10^{-3}$), followed by subsp. *cuspidata* (Nei, $H = 2.6 \times 10^{-3}$). Interestingly, subsp. *guanchica* is endemic to the Canary Islands, while the other two subspecies have a broad distribution, with subsp. *europaea*, in the Mediterranean Basin; and subsp. *cuspidata*, from South Africa to South Egypt and from Arabia to North India and South-West China. These results suggest that subsp. *guanchica* may originate from an ancient larger population with a broader ancestral distribution and was able to maintain their genetic diversity for a very long period.

Analysis of heterozygous SNVs

Polyploid subspecies show a high number of heterozygous SNVs (Fig. 1b, light striped colours), which may affect their phylogenetic positions because to reconstruct the nuclear phylogeny we only used homozygous positions (see Methods). To overcome this problem, we compared independently exclusive SNVs of each diploid subspecies (SNV positions only present in the analysed individual, polyploid individuals are excluded in this step) to those of the polyploids. Our results show that the polyploid subspecies share more exclusive homozygous SNVs with the subsp. *laperrinei* and *guanchica*, and more

exclusive heterozygous SNVs with the subsp. *laperrinei*, *cuspidata*, and *guanchica* (Fig. 3a). These results support the idea that at the nuclear level, subsp. *laperrinei* is involved in the origin of the polyploids.

Phylogenomic relationships in the *O. europaea* complex

To understand the evolutionary relationships among subspecies and sequenced individuals within the *O. europaea* complex, we ran a model-based genetic structure analysis using nuclear SNVs, and reconstructed phylogenetic trees using nuclear, mitochondrion, and plastid SNVs, separately (Fig. 4). Population structure analysis showed that the most likely number of ancestral genetic groups (k) among the *O. europaea* complex is $k=3$ (Fig. 4d). These three clusters of genetic ancestry are differentially present among the various individuals (Fig. 4d, Additional file 1: Table S4). Only one of the groups (cluster 3) is exclusively present in one of the subspecies (*cuspidata*), supporting a deep divergence of this subspecies from the rest of the *Olea* complex. Another genetic cluster (1) is particularly found in subsp. *europaea* and *guanchica*, but it is also present, to a lesser proportion, in three other subspecies (*cerasiformis*, *laperrinei*, *maroccana*). The remaining genetic cluster (2) is more abundant in *cerasiformis*, *laperrinei*, *maroccana*, and *cuspidata* from Iran. This latter individual shows a higher proportion of the genetic cluster 2, 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 relatively recent admixture among multiple founder populations, but they can also result from shared ancestry already present before the divergence of the populations [23, 24].

Additionally, we reconstructed the phylogenomic relationships of the *O. europaea* complex using plastid, mitochondrial, and nuclear SNVs. The three genomes will tell us different evolutionary histories and help to unravel reticulation or incomplete lineage sorting [7, 25, 26]. Since in *O. europaea*, plastid and mitochondria are maternally inherited [27], they are expected to reflect similar evolutionary events. However, the nuclear genome can show a more complex history due to admixture and recombination between intermixing populations. Indeed, our results show that in the plastid and mitochondrial trees (Fig. 4a,b) the individuals group according to the previously described plastid types [7, 11]. However, when we compare the plastid and mitochondrial trees to the nuclear tree, we observe some incongruences (Fig. 4a,b,c), which is taken in other studies as evidence of hybridization or admixture processes [28–30].

Importantly, the subsp. *laperrinei* has a complex hybridization history, which suggests a likely

Table 2 Nei's gene diversity (H) of the subsp. *europaea* var. *europaea*, subsp. *europaea* var. *sylvestris*, subsp. *guanchica*, and subsp. *cuspidata*

<i>Olea europaea</i> subsp.	H
<i>europaea</i> var. <i>europaea</i>	2.31E-03
<i>europaea</i> var. <i>sylvestris</i>	2.33E-03
<i>guanchica</i>	2.75E-03
<i>cuspidata</i>	2.56E-03

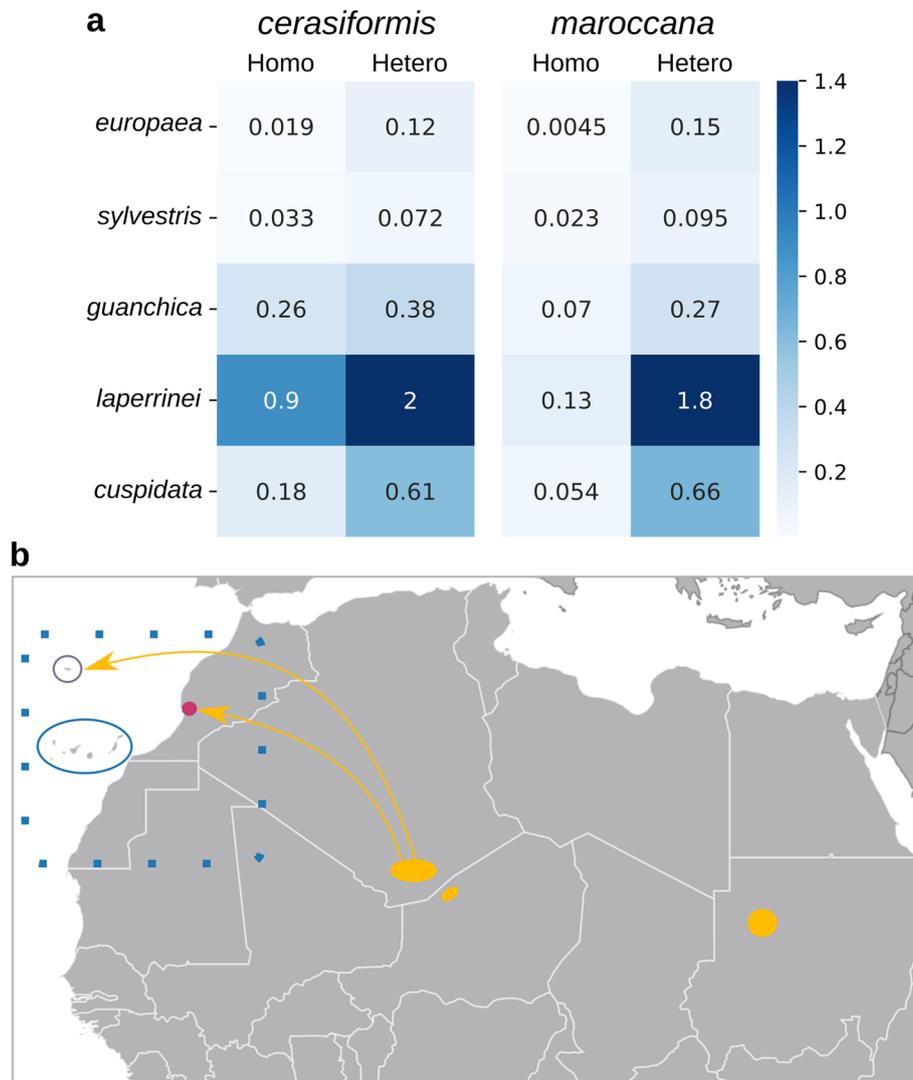


Fig. 3 The SNV similarity between diploids and polyploids and the possible origin of the polyploids. **a** Heatmap showing the percentage of the SNV positions of the polyploids (*cerasiformis* and *maroccana*) shared with the diploid subspecies. Only the exclusive SNVs of the diploid subspecies were analysed. **b** Map depicting the possible origin of the polyploids. Approximate distribution of subsp. *laperrinei* in yellow and subsp. *maroccana* in pink. The empty circles show the distribution of subsp. *cerasiformis* (purple) and *guanchica* (blue). The dotted line rectangle represents the presumed ancestral distribution of subsp. *guanchica*. The distribution of the subspecies was taken from [22]

involvement in the domestication of olive and the origin of polyploids. For instance, in the plastid tree, this subspecies appears closer to the subsp. *europaea* individuals that present eastern Mediterranean plastid type (E1) (Fig. 4a) and in the mitochondrial phylogeny, appears closer to the eastern and western type (E1, E2) (Fig. 4b), but in the nuclear tree, groups with the polyploid individuals (*cerasiformis* and *maroccana*) (Fig. 4c). These results indicate that hybridization occurred between *laperrinei* and *europaea*, which is in agreement with previous studies based on nuclear markers and plastid genomes [7, 31, 32]. In addition,

these results support the possibility that this subspecies is involved in the origin of the polyploids.

The subsp. *guanchica* also shows a reticulate evolution. This subspecies appears closer to the subsp. *europaea* in the nuclear tree, while groups with the polyploid individuals in the plastid and mitochondrial trees. These results indicate that subsp. *guanchica* is also involved in the origin of the polyploids.

Not all the accessions of subsp. *europaea* cluster together in the plastid and mitochondrial trees (Fig. 4a,b). In the plastid tree, we can observe the subsp. *europaea* forms three different clusters in agreement with their

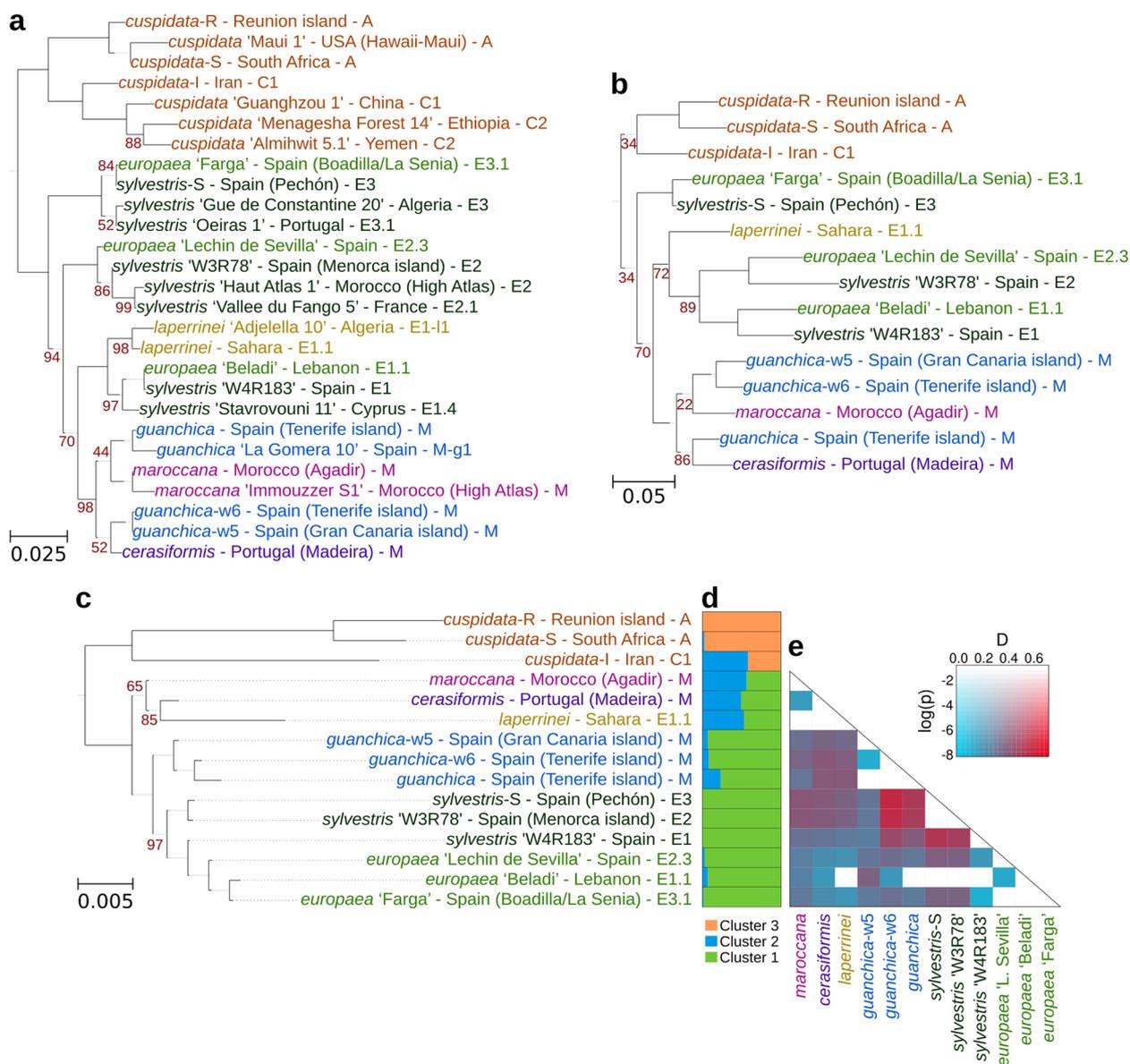


Fig. 4 Evolutionary relationships among the individuals of the *O. europaea* complex. Maximum likelihood species tree for the plastid SNVs (a), mitochondrial SNVs (b), and the nuclear SNVs (c). The geographical location of the accession and the plastid haplotype are indicated. Only bootstrap values below 100% are shown. d Bayesian clustering for the nuclear SNV data estimated in Structure v2.3 for the *O. europaea* complex. Structure bar plot shows the genetic clusters differentiated by colour. e Heatmap showing the D-statistic and its p value. Red colour indicates higher D-statistics, and more saturated colours indicate greater significance

plastid type, while in the mitochondrial tree two groups are formed. However, in the nuclear tree (Fig. 4c) all the individuals of the subsp. *europaea* group together, which indicates that they derived from a single ancestor(s), suggesting a monophyletic origin.

Phylogenetic analyses, and previously proposed phylogenetic relationships, suggest the existence of rampant gene flow in the evolution of the *O. europaea* complex [10]. A split network tree (Fig. 5) reveals a

heavily reticulated structure with conflicting phylogenetic signals affecting mostly the relationships among five subspecies, including the cultivated and wild olives, 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 [4, 8, 11]. The subspecies *europaea* and *guanchica* appear as monophyletic. Of note, the cultivars appear in an elongated, but also reticulated,

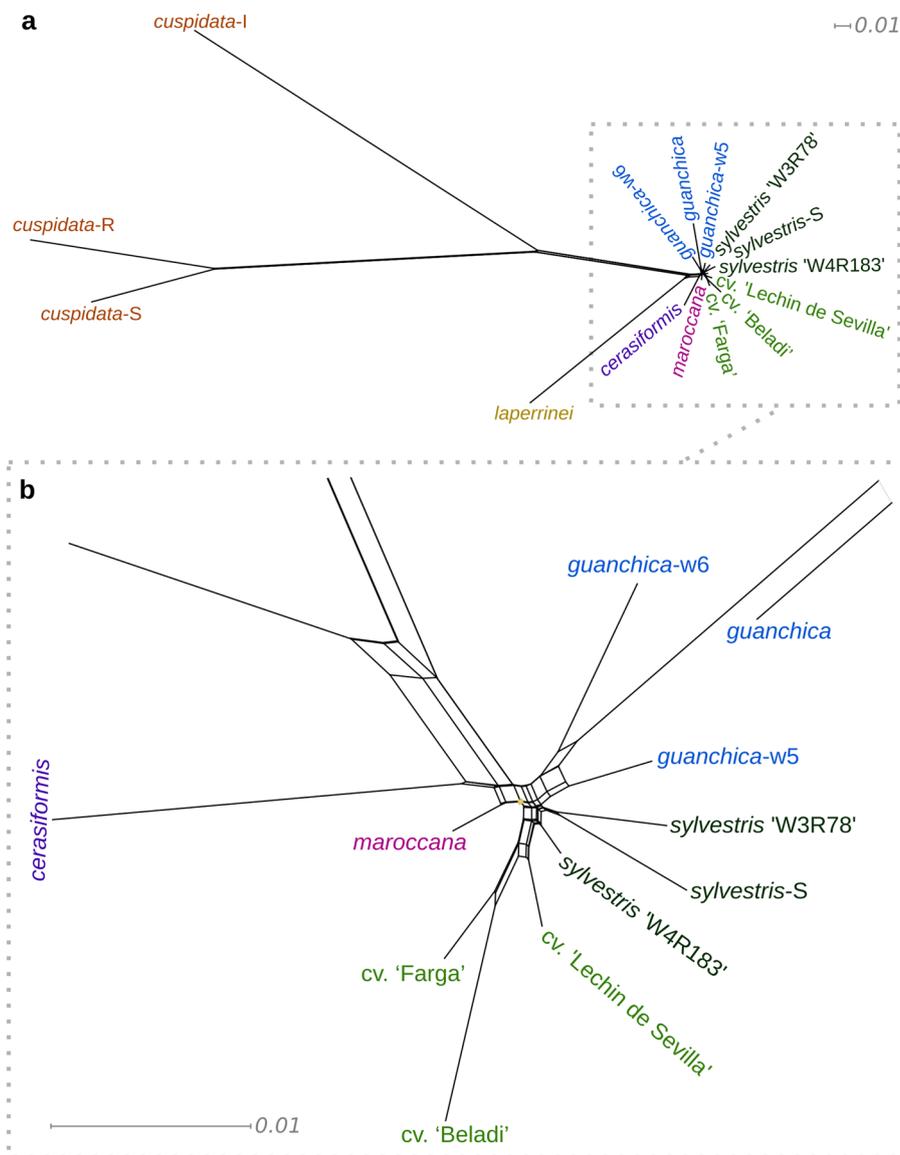


Fig. 5 SplitsTree derived from nuclear SNVs. **a** Complete plot. **b** Zoom in on the region in the rectangle. The neighbor-net method is used here to explore data conflict and not to estimate phylogeny

part of the network likely reflecting a more profound sequence divergence due to domestication.

The reticulated network can result from hybridization (including introgression) or incomplete lineage sorting. To search for evidence of introgression among our samples, we ran the ABBA-BABA test [33] using the program Dsuite v0.5 r49 [34] for all trios of the phylogenetic tree (Fig. 4c). This analysis shows that introgression occurred between subsp. *guanchica* and var. *sylvestris* (Fig. 4e). This result highlights the contribution of *guanchica* to subsp. *europaea*.

Discussion

The early-most olive divergence

Genomic differences between the individuals of subsp. *cuspidata* are remarkable. As mentioned above, individuals from Reunion Island and South Africa (*cuspidata-R* and *cuspidata-S*) group together in a single genetic group, while *cuspidata-I*, from Iran, is a mixture of two clusters (Fig. 4d). Indeed, *cuspidata-I* shares more genetic ancestry with other subspecies from North Africa than with the other *cuspidata* from Africa (Additional file 1: Table S4). Moreover, in the nuclear and the organellar trees, the individuals of subspecies *cuspidata* form

a monophyletic group (Fig. 4a,b,c). Interestingly, in all cases, the three individuals sequenced in this study are divided into two clades separated by a long branch: *cuspidata*-S and *cuspidata*-R on one side, and *cuspidata*-I on the other. Although our study includes a few individuals from subsp. *cuspidata*, our results show that African individuals highly differentiated from the Iranian individual. Differences between *cuspidata* from Africa and Asia were observed in previous studies based on organellar markers [7, 35–38]. These studies have shown that two geographically distant groups exhibit two different plastid types: “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 olives suggesting the possibility of sporadic hybridization among them [39–41]. However, for our sequenced individual, introgression from cultivars can hardly explain the observed pattern of genetic ancestry, as cluster 2 from our structure analysis is only residually present in cultivars, whereas is common in other wild *O. europaea* subspecies from Africa. Interestingly, a recent analysis of plant reproductive structures in Asian and African *cuspidata* accessions has shown numerous differences at morpho-structural and functional levels, despite this variability was suggested to be due to different adaptability to the growth environment [42], may also indicate genetic differences. Considering the phylogenetic divergence between the Asian (Iran) individual and the other *cuspidata* accessions, one more taxon should be considered in the *Olea europaea* complex: subsp. *ferruginea*.

Complex biogeographic scenario between the western Mediterranean and Macaronesian Islands

The present distribution of the *O. europaea* complex makes it an excellent model to study the colonisation of Macaronesian islands. Two subspecies, *guanchica* and *cerasiformis*, are endemic to the Macaronesian region and previous studies suggested two independent colonisation events to the Madeira and Canary Islands [43]. Our phylogenetic analysis shows that subsp. *guanchica* is genetically closely related to subsp. *europaea* (Fig. 4c). Moreover, despite being endemic to an oceanic archipelago (Canary Islands), its genetic diversity is higher than that within the other subspecies with wider distributions (subsp. *cuspidata*, subsp. *europaea*, Table 2). This high genetic diversity was also observed in other studies, which suggested that gene flow from subsp. *europaea* was the main driving factor [44]. We hypothesise that subsp. *guanchica* may originate at least once directly from the Mediterranean region. As the cultivated olive (subsp. *europaea*) has been planted across the Canary Islands since the arrival of Europeans, we cannot rule out gene flow with trees of the Canarias endemic subsp. *guanchica*.

On the other hand, the colonisation of subsp. *cerasiformis* to a Macaronesian Island (Madeira) is a more complex scenario, which involves a polyploidization event. Recent studies highlight the importance of polyploidy in the evolution of island biodiversity [45]. However, to understand the role of polyploids in islands, we need to understand their origin. Up to date, four main hypotheses have been postulated to explain the origin of the polyploid subspecies *cerasiformis* and the other continental endemic polyploid subspecies (*maroccana*). For each polyploid, we can find the following hypotheses: i) *cerasiformis* = *guanchica* x *europaea* [9], *cerasiformis* = ancestor of *guanchica* x *guanchica* [44] ii) *cerasiformis* = *laperrinei* x *laperrinei* (triploid-bridge) [22, 46], iii) *maroccana* = *laperrinei* x *laperrinei* (triploid-bridge) [22, 46], and iv) *maroccana* = *cerasiformis* x *guanchica* [44]. All these hypotheses have in common subsp. *guanchica* and *laperrinei* as parental taxa, albeit none of them has postulated the combination of both subspecies for the origin of the two polyploids. Our results suggest that ancestral populations of subsp. *guanchica* acted as a maternal donor (Fig. 4a,b), while ancestral populations of subsp. *laperrinei* could be the paternal donor (Fig. 4c). However, the current geographical distribution of subsp. *laperrinei* (Saharan massifs) and subsp. *guanchica* (Canary Islands) makes difficult the explanation of the hybrid origin of the endemic polyploids (subsp. *cerasiformis*—Madeira and subsp. *maroccana*—Agadir Mountains) (Fig. 3b) [2, 22, 47]. Interestingly, for subsp. *laperrinei*, earlier studies have shown that its small and highly isolated populations are able to maintain high genetic diversity by the existence of relict trees persisting for a very long time associated with asexual multiplication [31]. Also, ancient gene flow between subsp. *laperrinei* and subsp. *europaea* during favourable periods was proposed [7, 31, 32]. Thus, we suggest that gene flow between subsp. *laperrinei* and subsp. *guanchica* may have occurred during similar periods. These findings show evidence that hybridizations with change in ploidy level from an ancestor shared with subsp. *laperrinei*, as paternal donor, and *guanchica*, as maternal donor, may have brought about the two polyploid subspecies (Fig. 3b): *cerasiformis* = *guanchica* x *laperrinei*, *maroccana* = *cerasiformis* x *laperrinei*, or *maroccana* = *cerasiformis* x *guanchica*. In sum, we hypothesise that an ancestral already polyploid subsp. *cerasiformis* is the one that colonised Madeira island.

Current geographic distributions of the subspecies do not help to reconstruct the two polyploidization events. Ancient colonisation of Macaronesian islands by the tetraploid Madeiran olive (subsp. *cerasiformis*) and the diploid Canarian olive (subsp. *guanchica*) took place a long time ago (last two million years [9]), which makes difficult any reliable reconstruction. The same is true for

the hexaploid olive of Agadir mountains (subsp. *maroccana*) and the diploid Saharan olive (subsp. *laperrinei*) in northern Africa where desertification has been broadly expanded in the last 50,000 years [7]. Massive extinctions may explain why major difficulties in conciliation between current olive distributions and parental lineages to generate the two olive polyploids. The hypothesis of significant olive extinction because of the expansion of the Sahara desert in different periods of the Pleistocene-Holocene is supported by biogeographic patterns such as those of the Rand flora that illustrates east–west plant disjunctions in Africa [48] and Canarian and southern Asian disjunctions in the case of the pine trees [49]. Patterns of significant olive extinction have also been supported by phylogenomics of whole genomes to better explain even more ancient polyploidization in *Olea* [10].

Phylogenomic evidence for the origin of the olive tree

Many studies have pointed to the eastern Mediterranean Basin as the place where the first domestication event of olive occurred around 6,000 years ago [50, 51]. Recent studies have shown that this primary domestication was followed by numerous secondary events across Mediterranean countries, particularly involving genetic admixture with wild populations (var. *sylvestris*) from the western Mediterranean Basin [10]. However, few studies highlight the contribution of other subspecies to the evolution of olive. Our phylogenomic analysis shows that two subspecies are closely related to olive, subsp. *laperrinei* and *guanchica*. For the subsp. *laperrinei*, the organellar trees (Fig. 4a,b) show a closer relationship with the cultivars from the eastern Mediterranean Basin. This is in agreement with previous studies based on nuclear and plastid markers, which reveal historical hybridization between *laperrinei* and *europaea* during waning and waxing of African lineages as a result of climatic fluctuations during the Pleistocene [4, 8]. On the other hand, the nuclear phylogeny shows that the subsp. *europaea* is closely related to subsp. *guanchica*, which is considered a biogeographic connection between the Mediterranean and the Canarian olives. In sum, subsp. *laperrinei* and *guanchica* represent closely-related genetic pools to the Mediterranean olive (subsp. *europaea*) because of their genomic origin (ancestral gene flow with subsp. *laperrinei*, and colonisation of the Canary Islands bringing about subsp. *guanchica*).

Conclusions

Altogether, based on the presented data a phylogeny-informed grouping of the nuclear genetic diversity present in seven subspecies of *Olea europaea* would result in the following clades: i) *ferruginea* (Asia), ii) *cuspidata* (Africa); iii) *laperrinei* (Sahara), iv) *cerasiformis*

(Madeira), v) *maroccana* (Agadir Mountains, Morocco), vi) *guanchica* (Canary Islands), and vii) *europaea* (Mediterranean basin). The last one includes the domestication process, which included recurrent admixture between eastern Mediterranean domesticates and western Mediterranean wild populations.

The incongruences found between organellar and nuclear trees may reflect historical hybridization between main lineages of the *Olea europaea* complex. In particular, we propose a hybrid origin of the polyploid (allopolyploid) subspecies *cerasiformis* and *maroccana*. Previous results based on molecular phylogenetics agree with our phylogenomic hypothesis in which the diploid subsp. *laperrinei* and subsp. *guanchica* contributed to the formation of both allopolyploid subspecies.

It is important to note that northern Africa has suffered dramatic climatic changes determining the distribution of plant species in recent times. This can explain why some *Olea europaea* lineages are now restricted to narrow areas that may have been larger in the past. Indeed, given the current distribution of *laperrinei* (Saharan massifs) and *guanchica* (Canary Islands), and the presence of polyploids between these two clades in Madeira (*cerasiformis*) and Agadir Mountains (*maroccana*), we propose that the two clades encountered and formed hybrids in a presumably larger area spanning from the Atlas mountains to Madeira. As a result, the subspecies *maroccana* and *cerasiformis* appear to be allopolyploids that successfully adapted to the specific local conditions (narrow endemics).

In summary, our results show recurrent patterns of hybridization among most lineages of the *O. europaea* complex followed by lineage divergence and geographical isolation in agreement with previous molecular studies [4, 8, 36, 41]. Nevertheless, the use of 27 olive organellar and nuclear genomes gives us the opportunity to perform a more reliable reconstruction of geographic and phylogenetic origins of the Mediterranean olive tree (subsp. *europaea*) and relatives (six more subspecies).

Methods

Genome sequences

A total of 15 nuclear genome samples were analysed to represent the highest levels of genetic diversity in the *Olea europaea* complex. In particular, we sequenced the genome of five of the six subspecies of the *Olea europaea* complex [2]: two of *cuspidata*, and one of *cerasiformis*, *guanchica*, *maroccana*, and *laperrinei* (Table 1). DNA of each sample was extracted from leaf tissue as described in [6] and their genomes were sequenced using Illumina HiSeq 2000 paired-end technology to a sequencing depth ranging from 24 to 34 × as described in [6]. In addition to these six samples, we also included the publicly available

data of six individuals of the subspecies *europaea*, one subsp. *cuspidata*, and two *guanchica* (see Table 1) [6, 10, 52]. In addition, we downloaded the plastid genome of twelve individuals belonging to the different subspecies of *O. europaea* from the NCBI database and one mitochondrial genome (Table 1).

Detection of single nucleotide variants and small insertions/deletions

Sequence reads from each sample were mapped onto the nuclear, plastid, and mitochondrial reference genomes of *O. europaea* subsp. *europaea* cv. 'Farga' [10] using BWA 0.7.6a-r433 [53]. Single nucleotide variants (SNV) and small insertions/deletions (indels) were identified with GATK HaplotypeCaller v4.1.8.1 [54], setting ploidy according to the described ploidy level of the subspecies, i. e. hexaploid for *maroccana*, tetraploid for *cerasiformis*, and diploid for the other four subspecies. For SNVs we used the following filters: mapping quality (MQ>40), quality by depth (QD>2), Fisher strand bias (FS<60), mapping quality rank-sum test (MQRankSum>-12.5), read pos rank-sum test (ReadPosRankSum>-8), strand odds ratio (SOR<3), read depth of coverage (DP>=10). We also removed both, sites with missing alleles and sites with more than two alleles. For indels we used the following filters: quality by depth (QD>2), quality (QUAL>30.0), Fisher strand bias (FS<200). We also removed sites with missing alleles and sites supported only by one individual.

Ploidy estimation using the relative coverage of alternative alleles in heterozygous sites

To assess the ploidy of each sequenced individual we used the nuclear SNV data and plotted the relative coverage of alternative alleles in heterozygous sites. We considered only the 23 pseudo-chromosomes and each individual was analysed independently. For each heterozygous position, we computed the relative coverage of alternative alleles 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 a hexaploid five peaks around 0.17, 0.33, 0.5, 0.67, 0.83. Finally, all plots were generated using the Seaborn [55] python package.

Functional analysis of selected genomic regions

In order to determine whether the secondary peak observed at alternative allele frequencies of 0.25 is the product of differences in repetitive sequences among the *O. europaea* complex, we mapped those positions to the annotations of the repetitive regions and genes. We selected all the positions with frequencies ranging from

0.05 to 0.20 and we map their positions to the gff files of the repetitive regions and gene annotations of the Oe9 genome (<https://denovo.cnag.cat/olive>). Then, per each individual, we counted the number of positions that fall inside the regions of repetitive elements or genes. Finally, for the genes associated with those positions, we performed GO term enrichment analyses using FatiGO [56].

Exclusive SNV analysis

To assess putative parental lineages of polyploid individuals, we assessed which diploid groups of the *O. europaea* complex share more SNVs with the polyploid individuals. The groups studied are the three individuals of subspecies *guanchica*, *cuspidata*, and *europaea* (var. *europaea* and *sylvestris*) and one individual of subsp. *laperrinei*. Exclusive SNVs were homozygous only or homozygous and heterozygous positions in the studied diploid group (i.e. A/A or A/A, A/T for subsp. *guanchica*), and homozygous only in the other diploid groups (i.e. T/T for var. *europaea*, var. *sylvestris*, subsp. *cuspidata*, subsp. *laperrinei*). Then, we searched for those positions in the polyploid individuals and if they shared at least one allele with the studied group, we counted that position as shared. Finally, we calculated the percentage that those shared SNVs represented from the total SNVs in the polyploids (subsp. *cerasiformis* homozygous SNVs=2,502,578, subsp. *cerasiformis* heterozygous SNVs=6,567,558, subsp. *maroccana* homozygous SNVs=1,457,218, subsp. *maroccana* heterozygous SNVs=9,844,351).

Admixture mapping

Because of the large number of polymorphic positions in the nuclear genomes of the *O. europaea* complex, and computational limitations, we generated 10 subsets of 100,000 randomly chosen polymorphic positions without overlaps and analysed them in parallel. Then we identified population structure without a priori grouping assumptions, using the Structure software v2.3.4 [57]. 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 6, considering independent alleles and admixture of individuals. Simulations were repeated 5 times for each value of K. The optimal number of genetic clusters was determined using the ΔK method [58] using the software Structure Harvester [59]. Finally, the optimal K value was visualised with DISTRUCT v1.1 [60].

Phylogenetic analysis

Phylogenetic trees were reconstructed using SNVs data from nuclear, plastid, and mitochondrial genomes, separately. For the nuclear genome, we included only homozygous positions, excluding all positions with heterozygous

SNVs. In all cases, the genome sequence of the sequenced individuals was obtained by replacing the SNV positions in the respective reference genome, resulting in a pseudo alignment of all the considered genomes. For the plastid genomes, we included additional sequences by aligning our genomes with the genomes available in the databases (Table 1) using MAFFT v7.305b [61]. All these alignments were trimmed using trimAl v1.4 [62] with options -st 1 and -complementary, to remove all the non-informative positions. The final alignment had 3,632,494 variable positions for the nuclear genome, 328 for the plastid genome, and 3,245 for the mitochondrial genome. Phylogenetic trees were reconstructed from these alignments using RAxML v8.2.12 [63] and the GTR model. Support values were calculated based on 100 bootstrap repetitions. Additionally, for the nuclear data, we reconstructed a phylogenetic network using SplitsTree4 v4.17.1 and the NeighborNet approach [64].

For the subspecies *guanchica*, *cuspidata* and *europaea* (var. *europaea* and var. *sylvestris*), we estimated the Nei's gene diversity index [65] in order to assess the proportion of nucleotide variation in each subspecies.

Analysis of introgression

The ABBA-BABA test [33] was used to search for introgression among the *O. europaea* samples. Dsuite 0.5 r49 [34] was employed to calculate the D-statistic from nuclear SNV data for all subsets of trios that were compatible with the previously reconstructed phylogenetic tree (see above) using subsp. *cuspidata* as outgroup. For multiple hypotheses testing, we applied a false discovery rate correction to the p-values [66]. Then a heatmap showing the D-statistic and its p-value was plotted for all pairs of individuals using the plot_d.rb script (https://github.com/millanek/tutorials/blob/master/analysis_of_introgression_with_snp_data/src/plot_d.rb).

Abbreviations

subsp.	Subspecies (singular)
sub spp.	Subspecies (plural)
var.	Variety
ENA	European Nucleotide Archive
NCBI	National Center for Biotechnology Information
SNVs	Single nucleotide variants
indels	Insertions/Deletions
ATP	Adenosine triphosphate
GO	Gene Ontology
P25	SNV positions with relative coverage of alternative alleles ranging from 0.05 to 0.25
MQ	Mapping quality
QD	Quality by depth
FS	Fisher strand bias
MQRankSum	Mapping quality rank-sum test
ReadPosRankSum	Read pos rank-sum test
SOR	Strand odds ratio
DP	Depth of coverage

QUAL	Quality
MCMC	Markov chain Monte Carlo
GTR	General Time Reversible

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12915-023-01583-5>.

Additional file 1: Table S1. Number of small insertions/deletions (indels) per kilobase. The columns in order show the number of homozygous indels, heterozygous indels, total number of indels, large indels (>100 nucleotides), indels in genic regions. The *O. europaea* individuals are in rows. **Table S2.** Number of genes of each *O. europaea* individual containing indels and GO term enrichment analysis of those genes. **Table S3.** Analysis of the positions with relative coverage of alternative alleles in heterozygous sites values ranging from 0.05 to 0.25 (positions P25). The columns show: name of the individual, number of positions P25, number of P25 positions present in repetitive regions, percentage of P25 positions present in repetitive regions, number of P25 positions present in genes, percentage of P25 positions present in genes, number of genes associated with P25 positions, GO term enriched function of the genes associated with P25 positions with *p*-value < 0.01. **Table S4.** Admixture coefficient (Q) of each individual per cluster. This table was used to create Fig. 4d.

Acknowledgements

The authors want to thank Marina Marcet-Houben, Fernando Cruz, and Tyler Alyoto for their help with data access and data submission.

Authors' contributions

IJ performed bioinformatics analysis. IJ and TG drafted the first version of the manuscript to which all authors contributed. PV provided samples, designed the selection of individuals, and provided data and important scientific insights to interpret the results. TG supervised the study. All the authors read and approved the final manuscript.

Authors' information

Irene Julca: Twitter (@IreneJulca).
Pablo Vargas: Twitter (@pablovargas1111).
Toni Gabaldon: Twitter (@toni_gabaldon).

Funding

This research has received funding from the Spanish Ministry of Science and Innovation for grant PGC2018-099921-B-I00, cofounded by European Regional Development Fund (ERDF); from the Catalan Research Agency (AGAUR) SGR423; from the European Union's Horizon 2020 research and innovation programme (ERC-2016–724173); from the Gordon and Betty Moore Foundation (Grant GBMF9742) and from the Instituto de Salud Carlos III (INB Grant PT17/0009/0023—ISCIII-SGEFI/ERDF). IJ is supported by the Singaporean Ministry of Education grant MOE2018-T2-2–053. TG and PV acknowledge support from Banco Santander for the Olive Genome Sequencing project.

Availability of data and materials

All data generated or analysed during this study are included in this published article, its additional files and publicly available repositories. The sequencing data produced in this project have been deposited in the ENA (European Nucleotide Archive) under the accession number PRJNA841802 [67]. The other datasets analysed in this study are available in the NCBI and ENA databases, and the accession numbers are indicated in Table 1.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Barcelona Supercomputing Centre (BSC-CNS), Plaça Eusebi Güell, 1-3, 08034 Barcelona, Spain. ²Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain. ³School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551, Singapore. ⁴Department of Biodiversity and Conservation, Real Jardín Botánico de Madrid, Calle Claudio Moyano 1, 28014 Madrid, Spain. ⁵Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain. ⁶CIBER de Enfermedades Infecciosas, Instituto de Salud Carlos III, Madrid, Spain.

Received: 12 January 2023 Accepted: 30 March 2023

Published online: 17 April 2023

References

- Green PS, Wickens GE. The *Olea europaea* complex. Edinburgh: The Davis & Hedge Festschrift, Edinburgh University Press; 1989.
- Green PS. A revision of *Olea* L. (Oleaceae). *Kew Bull.* 2002;57:91.
- Vargas P, Muñoz Garmendia F, Hess J, Kadereit J. *Olea europaea* subsp. *ganchica* and subsp. *maroccana* (Oleaceae), two new names for olive tree relatives. *Anales del Jardín Botánico de Madrid.* 2000;58:360–1.
- Besnard G, Rubio de Casas R, Christin P-A, Vargas P. Phylogenetics of *Olea* (Oleaceae) based on plastid and nuclear ribosomal DNA sequences: tertiary climatic shifts and lineage differentiation times. *Ann Bot.* 2009;104:143–60.
- Julca I, Marcet-Houben M, Vargas P, Gabaldón T. Phylogenomics of the olive tree (*Olea europaea*) reveals the relative contribution of ancient allo- and autopolyploidization events. *BMC Biol.* 2018;16:15.
- Cruz F, Julca I, Gómez-Garrido J, Loska D, Marcet-Houben M, Cano E, et al. Genome sequence of the olive tree. *Olea europaea* Gigascience. 2016;5:29.
- Besnard G, Rubio de Casas R, Vargas P. Plastid and nuclear DNA polymorphism reveals historical processes of isolation and reticulation in the olive tree complex (*Olea europaea*). *J Biogeography.* 2007;34:736–52.
- Rubio de Casas R, Besnard G, Schönswetter P, Balaguer L, Vargas P. Extensive gene flow blurs phylogeographic but not phylogenetic signal in *Olea europaea* L. *Theor Appl Genet.* 2006;113:575–83.
- Besnard G, García-Verdugo C, De Casas RR, Treier UA, Galland N, Vargas P. Polyploidy in the olive complex (*Olea europaea*): evidence from flow cytometry and nuclear microsatellite analyses. *Ann Bot.* 2008;101:25–30.
- Julca I, Marcet-Houben M, Cruz F, Gómez-Garrido J, Gaut BS, Díez CM, et al. Genomic evidence for recurrent genetic admixture during the domestication of Mediterranean olive trees (*Olea europaea* L.). *BMC Biol.* 2020;18:148.
- Besnard G, Hernández P, Khadari B, Dorado G, Savolainen V. Genomic profiling of plastid DNA variation in the Mediterranean olive tree. *BMC Plant Biol.* 2011;11:80.
- Hanada K. Serine palmitoyltransferase, a key enzyme of sphingolipid metabolism. *Biochim Biophys Acta.* 2003;1632:16–30.
- Inês C, Parra-Lobato MC, Paredes MA, Labrador J, Gallardo M, Saucedo-García M, et al. Sphingolipid Distribution, Content and Gene Expression during Olive-Fruit Development and Ripening. *Front Plant Sci.* 2018;9:28.
- Corbacho J, Inês C, Paredes MA, Labrador J, Cordeiro AM, Gallardo M, et al. Modulation of sphingolipid long-chain base composition and gene expression during early olive-fruit development, and putative role of brassinosteroid. *J Plant Physiol.* 2018;231:383–92.
- Noormohammadi Z, Samadi-Molayousefi H, Sheidai M. Intra-specific genetic diversity in wild olives (*Olea europaea* ssp *cuspidata*) in Hormozgan Province. *Iran Genet Mol Res.* 2012;11:707–16.
- Hwang J-U, Song W-Y, Hong D, Ko D, Yamaoka Y, Jang S, et al. Plant ABC transporters enable many unique aspects of a terrestrial plant's lifestyle. *Mol Plant.* 2016;9:338–55.
- Mascagni F, Barghini E, Giordani T, Rieseberg LH, Cavallini A, Natali L. Repetitive DNA and Plant Domestication: Variation in Copy Number and Proximity to Genes of LTR-Retrotransposons among Wild and Cultivated Sunflower (*Helianthus annuus*) Genotypes. *Genome Biol Evol.* 2015;7:3368–82.
- Banasiak J, Jasiński M. ATP-binding cassette transporters in nonmodel plants. *New Phytol.* 2022;233:1597–612.
- Stuessy TF, Takayama K, López-Sepúlveda P, Crawford DJ. Interpretation of patterns of genetic variation in endemic plant species of oceanic islands. *Bot J Linn Soc.* 2014;174:276–88.
- Montgomery ME, Woodworth LM, Nurthen RK, Gilligan DM, Briscoe DA, Frankham R. Relationships between population size and loss of genetic diversity: comparisons of experimental results with theoretical predictions. *Conserv Genet.* 2000;1:33–43.
- Díez CM, Trujillo I, Martínez-Urdiroz N, Barranco D, Rallo L, Marfil P, et al. Olive domestication and diversification in the Mediterranean Basin. *New Phytol.* 2015;206:436–47.
- Besnard G, Anthelme F, Baali-Cherif D. The Laperrine's olive tree (Oleaceae): a wild genetic resource of the cultivated olive and a model-species for studying the biogeography of the Saharan Mountains. *Acta Botanica Gallica.* 2012;159:319–28.
- Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science.* 2008;319:1100–4.
- Mousavi S, Mariotti R, Bagnoli F, Costantini L, Cultrera NGM, Arzani K, et al. The eastern part of the Fertile Crescent concealed an unexpected route of olive (*Olea europaea* L.) differentiation. *Ann Bot.* 2017;119:1305–18.
- Linder CR, Rieseberg LH. Reconstructing patterns of reticulate evolution in plants. *Am J Bot.* 2004;91:1700–8.
- Petit RJ, Duminil J, Fineschi S, Hampe A, Salvini D, Vendramin GG. Comparative organization of chloroplast, mitochondrial and nuclear diversity in plant populations. *Mol Ecol.* 2005;14:689–701.
- Besnard G, Khadari B, Villemur P, Bervillé A. Cytoplasmic male sterility in the olive (*Olea europaea* L.). *Theor Appl Genet.* 2000;100:1018–24.
- Fehrer J, Gemeinholzer B, Chrtek J, Bräutigam S. Incongruent plastid and nuclear DNA phylogenies reveal ancient intergeneric hybridization in *Pilosella* hawkweeds (Hieracium, Cichorieae, Asteraceae). *Mol Phylogenet Evol.* 2007;42:347–61.
- Barber JC, Finch CC, Francisco-Ortega J, Santos-Guerra A, Jansen RK. Hybridization in Macaronesian *Sideritis* (Lamiaceae): Evidence from incongruence of multiple independent nuclear and chloroplast sequence datasets. *Taxon.* 2007;56:74–88.
- Pelser PB, Kennedy AH, Tepe EJ, Shidler JB, Nordenstam B, Kadereit JW, et al. Patterns and causes of incongruence between plastid and nuclear *Senecioneae* (Asteraceae) phylogenies. *Am J Bot.* 2010;97:856–73.
- Baali-Cherif D, Besnard G. High genetic diversity and clonal growth in relict populations of *Olea europaea* subsp. *laperrinei* (Oleaceae) from Hoggar, Algeria. *Ann Bot.* 2005;96:823–30.
- Besnard G, El Bakkali A, Haouane H, Baali-Cherif D, Moukhlil A, Khadari B. Population genetics of Mediterranean and Saharan olives: geographic patterns of differentiation and evidence for early generations of admixture. *Ann Bot.* 2013;112:1293–302.
- Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. *Science.* 2010;328:710–22.
- Malinsky M, Matschiner M, Svárdal H. Dsuite - Fast D-statistics and related admixture evidence from VCF files. *Mol Ecol Resour.* 2021;21:584–95.
- Besnard G, Khadari B, Baradat P, Bervillé A. *Olea europaea* (Oleaceae) phylogeography based on chloroplast DNA polymorphism. *Theor Appl Genet.* 2002;104:1353–61.
- Besnard G, Khadari B, Baradat P, Bervillé A. Combination of chloroplast and mitochondrial DNA polymorphisms to study cytoplasm genetic differentiation in the olive complex (*Olea europaea* L.). *Theor Appl Genet.* 2002;105:139–44.
- Besnard G, Baradat P, Breton C, Khadari B, Bervillé A. Olive domestication from structure of oleasters and cultivars using nuclear RAPDs and mitochondrial RFLPs. *Genet Sel Evol.* 2001;33:5251–68.
- Lumaret R, Amame M, Ouazzani N, Baldoni L, Debain C. Chloroplast DNA variation in the cultivated and wild olive taxa of the genus *Olea* L. *Theor Appl Genet.* 2000;101:547–53.
- Sheidai M, Noormohammadi Z, Dehghani A, Parvini F, Hoshjar-Parsian H, Hosseini-Mazinani M. Intra-specific morphological and molecular diversity in brown olive (*Olea cuspidata*) of Iran. *ScienceAsia.* 2010;36:187–93.

40. Hosseini-Mazinani M, Mariotti R, Torkzaban B, Sheikh-Hassani M, Ataei S, Cultrera NGM, et al. High genetic diversity detected in olives beyond the boundaries of the Mediterranean Sea. *PLoS ONE*. 2014;9: e93146.
41. Besnard G, El Bakkali A. Sequence analysis of single-copy genes in two wild olive subspecies: nucleotide diversity and potential use for testing admixture. *Genome*. 2014;57:145–53.
42. Caceres ME, Pupilli F, Sarri V, Mencuccini M, Ceccarelli M. Floral biology in *Olea europaea* subsp. *cuspidata*: A comparative structural and functional characterization Flora-Morphology. *Distribution, Functional Ecology of Plants*. 2016;222:27–36.
43. Hess J, Kadereit JW, Vargas P. 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). *Mol Ecol*. 2000;9:857–68.
44. García-Verdugo C, Fay MF, Granada-Yela C, Casas DE, RR, Balaguer L, Besnard G, et al. Genetic diversity and differentiation processes in the ploidy series of *Olea europaea* L.: a multiscale approach from subspecies to insular populations. *Mol Ecol*. 2009;18:454–67.
45. Meudt HM, Albach DC, Tanentzap AJ, Igea J, Newmarch SC, Brandt AJ, et al. Polyploidy on islands: its emergence and importance for diversification. *Front Plant Sci*. 2021;12: 637214.
46. Besnard G, Baali-Cherif D. Coexistence of diploids and triploids in a Saharan relict olive: evidence from nuclear microsatellite and flow cytometry analyses. *C R Biol*. 2009;332:1115–20.
47. Médail F. 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. *Botan J Linn Soc*. 2001;137:249–66.
48. Sanmartín I, Anderson CL, Alarcon M, Ronquist F, Aldasoro JJ. Bayesmartin island biogeography in a continental setting: the Rand Flora case. *Biol Lett*. 2010;6:703–7.
49. Jin W-T, Gernandt DS, Wehenkel C, Xia X-M, Wei X-X, Wang XQ. Phylogenomic and ecological analyses reveal the spatiotemporal evolution of global pines. *Proc Natl Acad Sci USA*. 2021;118.
50. Besnard G, Khadari B, Navascués M, Fernández-Mazuecos M, El Bakkali A, Arrigo N, et al. The complex history of the olive tree: from Late Quaternary diversification of Mediterranean lineages to primary domestication in the northern Levant. *Proc Biol Sci*. 2013;280:20122833.
51. Besnard G, Terral J-F, Cornille A. On the origins and domestication of the olive: a review and perspectives. *Ann Bot*. 2018;121:385–403.
52. Gros-Balthazard M, Besnard G, Sarah G, Holtz Y, Leclercq J, Santoni S, et al. Evolutionary transcriptomics reveals the origins of olives and the genomic changes associated with their domestication. *Plant J*. 2019;100:143–57.
53. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25:1754–60.
54. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20:1297–303.
55. Waskom M, Botvinnik O, O’Kane D, Hobson P, Lukauskas S, Gemperline DC, et al. Mwaskom/Seaborn: V0.8.1 (September 2017). Zenodo. 2017. Available from: <https://doi.org/10.5281/zenodo.883859>.
56. Al-Shahrour F, Minguez P, Tárraga J, Medina I, Alloza E, Montaner D, et al. FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic Acids Res*. 2007;35(Web Server issue):W91–6.
57. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*. 2000;155:945–59.
58. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol*. 2005;14:2611–20.
59. Earl DA, vonHoldt BM. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour*. 2012;4:359–61.
60. Rosenberg NA. DISTRUCT: a program for the graphical display of population structure. *Mol Ecol Notes*. 2004;4:137–8.
61. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*. 2005;33:511–8.
62. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 2009;25:1972–3.
63. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30:1312–3.
64. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol*. 2006;23:254–67.
65. Nei M. Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci USA*. 1973;70:3321–3.
66. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc: Ser B (Methodol)*. 1995;57:289–300.
67. Julca I, Marcet-Houben M, Vargas P, Gabaldón T. Whole genome shotgun sequencing of six individuals of *Olea europaea*. GenBank. 2023.

Publisher’s Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

