the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2023) doi: 10.1111/tpj.16270

RESOURCE

Whole genome scanning of a Mediterranean basin hotspot collection provides new insights into olive tree biodiversity and biology

Christos Bazakos[1,2,3,*], Konstantinos G. Alexiou[4,5], Sebastián Ramos-Onsins[4,5], Georgios Koubouris[6], Nikolaos Tourvas[7], Aliki Xanthopoulou[1,2], Ifigeneia Mellidou[1,2], Theodoros Moysiadis[1,8], Ioanna-Theoni Vourlaki[4,5], Ioannis Metzidakis[6], Chrysi Sergentani[6], Ioanna Manolikaki[6], Michail Michailidis[9], Adamantia Pistikoudi[10], Alexios Polidoros[10], George Kostelenos[11], Filippos Aravanopoulos[7], Athanassios Molassiotis[9] and Ioannis Ganopoulos[1,2,*]

[1]Institute of Plant Breeding and Genetic Resources, Hellenic Agricultural Organization (ELGO) DIMITRA, Thermi 57001, Thessaloniki, Greece,

Joint Laboratory of Horticulture, Hellenic Agricultural Organization (ELGO) DIMITRA, Thermi 57001, Thessaloniki, Greece,

Department of Comparative Development and Genetics, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10 50829, Cologne, Germany,

Centre for Research in Agricultural Genomics, CSIC-IRTA-UAB-UB, Campus UAB, Barcelona, Spain,

Institut de Recerca i Tecnologia Agroalimentàries, Barcelona, Spain,

Institute of Olive Tree, Subtropical Crops and Viticulture, Hellenic Agricultural Organization (ELGO) DIMITRA, Chania 73134, Greece,

Laboratory of Forest Genetics, Faculty of Agriculture, Forestry and Natural Environment, Aristotle University of Thessaloniki 54124, Thessaloniki, Greece,

Department of Computer Science, School of Sciences and Engineering, University of Nicosia, Nicosia 2417, Cyprus,

Laboratory of Pomology, Department of Horticulture, Aristotle University of Thessaloniki, Thermi 57001, Thessaloniki, Greece,

Laboratory of Genetics and Plant Breeding, School of Agriculture, Aristotle University of Thessaloniki 54124, Thessaloniki, Greece, and

Kostelenos Olive Nurseries, 18020, Poros-Trizinias, Greece

Received 25 January 2023; revised 29 March 2023; accepted 29 April 2023.
*For correspondence (e-mail cbazakos@elgo.gr; giannis.ganopoulos@gmail.com).

SUMMARY

Olive tree (*Olea europaea* L. subsp. *europaea* var. *europaea*) is one of the most important species of the Mediterranean region and one of the most ancient species domesticated. The availability of whole genome assemblies and annotations of olive tree cultivars and oleaster (*O. europaea* subsp. *europaea* var. *sylvestris*) has contributed to a better understanding of genetic and genomic differences between olive tree cultivars. However, compared to other plant species there is still a lack of genomic resources for olive tree populations that span the entire Mediterranean region. In the present study we developed the most complete genomic variation map and the most comprehensive catalog/resource of molecular variation to date for 89 olive tree genotypes originating from the entire Mediterranean basin, revealing the genetic diversity of this commercially significant crop tree and explaining the divergence/similarity among different variants. Additionally, the monumental ancient tree 'Throuba Naxos' was studied to characterize the potential origin or routes of olive tree domestication. Several candidate genes known to be associated with key agronomic traits, including olive oil quality and fruit yield, were uncovered by a selective sweep scan to be under selection pressure on all olive tree chromosomes. To further exploit the genomic and phenotypic resources obtained from the current work, genome-wide association analyses were performed for 23 morphological and two agronomic traits. Significant associations were detected for eight traits that provide valuable candidates for fruit tree breeding and for deeper understanding of olive tree biology.

Keywords: comparative genomics, genomic resources, GWAS, olive tree domestication, WGRS.

INTRODUCTION

Cultivated olive tree (*Olea europaea* L. subsp. *europaea* var. *europaea*) has been the most emblematic perennial crop in the Mediterranean region, and its origin is found in ancient civilizations dating back six millennia (Loumou & Giourga, 2003). Olive tree cultivation accounts for a significant part of the agro-industrial sector in Mediterranean areas (Vossen, 2007); it not only represents a considerable economic and agricultural factor, but also plays a role in the maintenance of biodiversity, showing a link with some wild crop relatives. The wild olive tree (*O. europaea* subsp. *europaea* var. *sylvestris*), also known as oleaster, which came from Asia Minor and then expanded to Greece, is considered to be the ancestor of cultivated olive varieties (Bartolini & Petruccelli, 2002). Olive tree cultivars have been planted nearby wild populations since ancient times, where they exchange pollen, which has resulted in effective crop production and historical hybridization (Rubio De Casas et al., 2006). Moreover, olive tree research and breeding have been limited by factors such as the time-consuming process of constructing cross populations for genetic mapping, the long juvenile phase, and the high levels of heterozygosity (Rugini et al., 2016). Therefore, the recovery of genetic diversity and the identification of genetic regions of olive tree associated with important agronomic traits linked to phenology, yield, and quality of the oil have become fundamental, notably in the context of the growing environmental impact of climate change (Aydin et al., 2021; Skodra et al., 2021).

The availability of whole genome assemblies and annotations of olive tree cultivars and oleasters (Barghini et al., 2014; Cruz et al., 2016; Jiménez-Ruiz et al., 2020; Rao et al., 2021; Unver et al., 2017) and the discovery of high-throughput resequencing technologies (Kersey, 2019) have contributed to a better understanding of genetic and genomic differences among olive tree cultivars (Aydin et al., 2021; Vatansever et al., 2022). The olive tree is a complex example of the domestication of fruit trees (Besnard et al., 2018); particularly the origins and genetic background of domestication-related phenotypic changes remain debatable (Gros-Balthazard et al., 2019). Several efforts have been made to pinpoint the molecular variations related to olive tree domestication and elucidate the evolutionary history of olive tree, for instance by resequencing 56 nuclear genomes (Julca et al., 2020), by investigating the transcriptomic data of 68 genotypes (Gros-Balthazard et al., 2019), and by analyzing the transposable elements of 51 resequenced genomes (Jiménez-Ruiz et al., 2020). The comprehensive investigation of the 68 olive tree transcriptomes (Gros-Balthazard et al., 2019) revealed that olive tree's domestication traits originated primarily through changes in gene expression, which is consistent with its evolutionary history. According to Julca et al. (2020), extensive hybridization has shaped the evolutionary history of several olive tree lineages. After the initial domestication in the eastern Mediterranean basin, numerous secondary events occurred in the majority of southern European and northern African countries, frequently involving admixture with genetically diverse wild populations, especially from the western Mediterranean basin (Julca et al., 2020). Contrariwise, Jiménez-Ruiz et al. (2020) performed genomic population analysis and supported two independent events in olive tree domestication, including an early putative genetic bottleneck.

To gain a better understanding of the patterns of genome-wide variation in olive tree, in this work we analyzed the genomic data of 89 olive tree genotypes. This panel includes 44 newly resequenced genotypes (41 cultivars, one monumental olive tree, and two genotypes of *O. europaea* subsp. *cuspidata* that were used as outgroup) in high depth from the National Olive Germplasm Bank of Greece (NOGB) and 45 genotypes from a previous study (Jiménez-Ruiz et al., 2020). The 40 Greek cultivars span the traditional range of olive tree cultivation and represent the majority of variation (>95%) in Greece (Koubouris et al., 2019; Xanthopoulou et al., 2014). Sequence analysis focused on genomic regions associated with propitious variation, such as deletions, substitutions, and duplications, providing the first comprehensive catalog of molecular variation in this species, which is helpful for explaining divergence/similarity among different variants. Furthermore, genealogical analyses of the 89 olive tree genotypes were conducted to investigate the phylogeny and migration history of cultivated and wild olive tree groups and to discern how Greek cultivars relate to other domesticated and wild populations. To this aim, the monumental tree 'Throuba Naxos' was included to characterize the role of Greek cultivars in the potential origin and/or routes of olive tree domestication.

The generation of high-coverage sequence data provides high-density single nucleotide polymorphism (SNP) markers across the genome, which significantly supports the mapping of genomic variants that are associated with morphological and agronomic traits (Bazakos et al., 2017). In several crops, high-density SNP markers have been used for the identification of domestication genes, quantitative trait locus (QTL) mapping, genome-wide association studies (GWAS), and population genomic screens for signatures of selection (Meyer & Purugganan, 2013). Previously, Gros-Balthazard et al. (2019) identified differentially expressed genes and screened the transcriptome for signatures of selection in 39 cultivated olive tree accessions and 27 oleasters. Herein, we took advantage of the large population and the high-density SNP markers across the genome to uncover by a selective sweep scan several candidate genes under selection pressure on all olive tree chromosomes. We also performed a GWAS to identify

genetic loci and candidate genes that may be linked to phenotypic variation in key agronomic traits, such as fruit weight, olive oil content, and ripening time, which are important for olive tree breeding.

RESULTS AND DISCUSSION

A high-density olive tree genomic variation map

High-quality whole genome resequencing (WGRS) data were generated for 89 olive tree genotypes including 76 cultivars originating from the Mediterranean basin (from Morocco to Syria), one Greek monumental olive tree, 10 genotypes from the wild species *O. europaea* subsp. *europaea* var. *sylvestris*, and two *O. europaea* subsp. *cuspidata* genotypes as outgroup (Table S1). Our dataset combined new sequencing data from 44 genomes with 45 publicly available whole genomes from cultivated and wild olive trees (Figure 1a; Table S1). WGRS analysis of the 89 unique olive tree genotypes (Figure 1a) generated a total of 1748 Gbp of sequences, with an average depth of 28× (Table S3).

Mapping results were used to identify small-scale variations present as SNPs and InDels (<50 bp), as well as larger-scale variants including genome-wide copy number variations (CNVs) and genic deletions (≥50 bp) and duplications (Table 1; Table S4). In total, a final set of 1 607 302 high-quality SNPs were identified, resulting in an average of one SNP every 356 bp. A total of 52 184 SNPs were found in coding regions, including 30 582 non-synonymous and 20 887 synonymous SNPs (Table S5). To check whether our variant calling from Illumina reads is reliable for downstream analyses, we sequenced 16 SNPs in 16 olive tree genotypes by Sanger sequencing, and most of the identified SNPs (98.05%) were validated (Table S6).

The analysis of InDels revealed a total of 3 313 827 variants, of which 1 384 758 were insertions and 1 929 069 were deletions. Among those with a potential functional impact, 1.44% were located within gene exons, and 0.2% were located in splice site regions, while 38.41% of InDels were found in 5′- and 3′-untranslated regions. The vast majority of the InDels was located upstream or downstream of genes, as well as in intergenic regions. The size

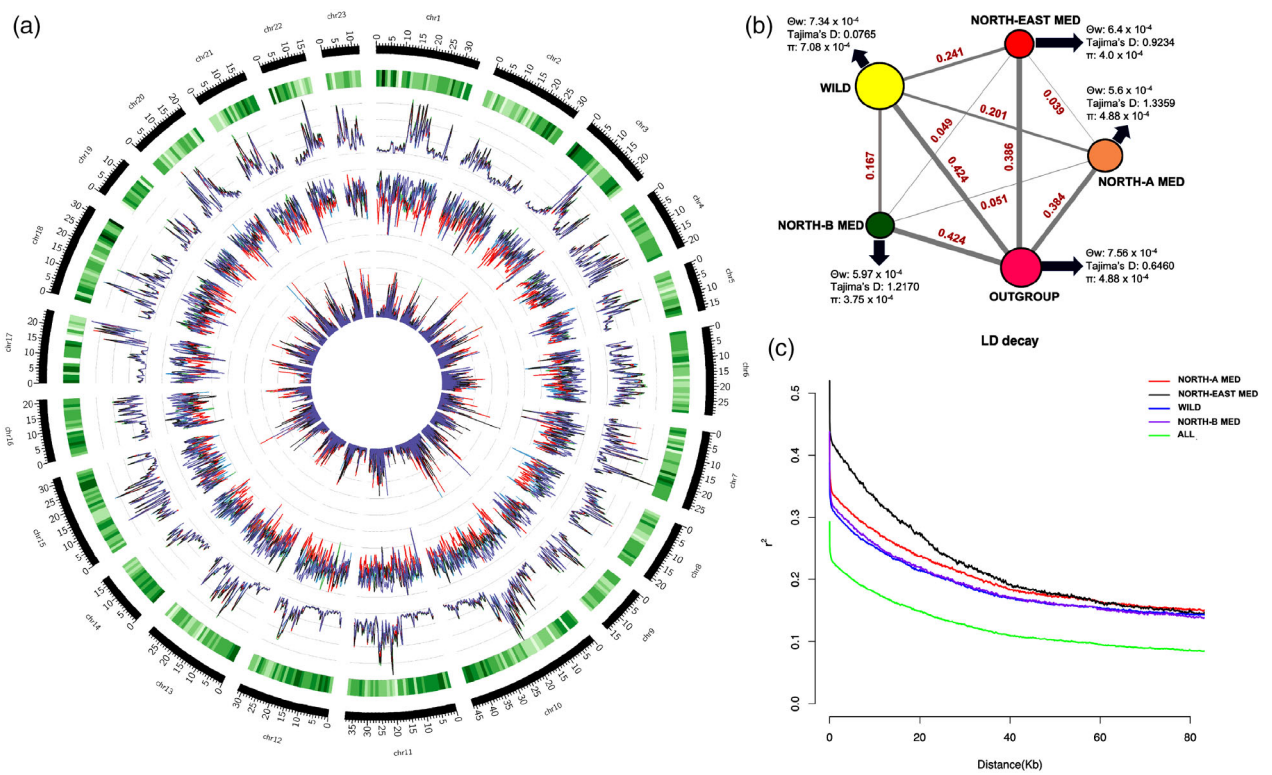

Figure 1. Genetic diversity and linkage disequilibrium (LD) for different olive tree groups.
(a) Circos plot for the genome-wide nucleotide diversity. Cycles from outside inwards: 1, repeat density; 2, gene density; the genome-wide distribution for θw, Tajima's *D*, and π values are shown in cycles. The line colors for each population are green (NORTH-A MED), blue (NORTH-EAST MED), dark red (WILD), very dark gray (NORTH-B MED), and dark purple (OUTGROUP).
(b) Summary of nucleotide diversity (π, θw, and Tajima's *D*) and population divergence (F_{ST}) across the five pre-defined groups. The value in each circle represents nucleotide diversity for the group, and values between pairs indicate population divergence (F_{ST}).
(c) LD decay between *Olea europaea* and various populations.

The Plant Journal published by Society for Experimental Biology and John Wiley & Sons Ltd.,
The Plant Journal, (2023), doi: 10.1111/tpj.16270

Table 1 Genome-wide variations identified in 89 olive genotypes

Variant	Type	Cultivars		Wild	Cultivars	Outgroup	Genotypes
		NORTH-A MED	NORTH-EAST MED	WILD	NORTH-B MED	*Olea europaea* subsp. *cuspidata*	
	Accessions (*n*)	33	14	11	29	2	89
SNPs	Total	1 124 097	75 269	1 280 605	871 411	262 239	1 607 302
	Intergenic	1 008 793	671 555	1 150 340	787 883	237 947	1 448 055
	Introns	76 073	53 488	85 114	54 625	15 018	103 601
	Exons	3674	25 865	42 321	27 155	8797	52184
InDels	Total	2 424 960	1 807 875	2 319 369	2 368 256	801 717	3 313 827
	Insertions	1 035 065	760 423	978 584	1 004 191	333 799	1 384 758
	Deletions (<50 bp)	1 389 895	1 047 452	1 340 785	1 364 065	467 918	1 929 069
Genome-wide CNVs	Total	1022	902	844	878	207	1905
	CNG	480	353	290	457	65	809
	CNL	542	549	554	421	142	1096
Genic and perigenic SVs	Total	840	687	692	756	52	913
	Deletions (≥50 bp)	783	637	647	707	50	851
	Duplications	57	50	45	49	2	62

CNG, copy number gain; CNL, copy number loss; CNV, copy number variant; SV, structural variant.

of deletions ranged from one to eight nucleotides in length. Most of the insertions and deletions (20.63%) consisted of a single nucleotide. Di- and trinucleotide InDels accounted for 7.53 and 5.85% of the total number of InDels, respectively (Table 1; Table S7.1). Similar results have been reported for the perennial sweet cherry tree species (*Prunus avium*) when the whole genomes of 21 sweet cherry accessions originating from the Mediterranean region were resequenced (Xanthopoulou et al., 2020).

Genome-wide CNV analysis revealed the presence of 809 copy number gains (CNGs) and 1096 copy number losses (CNLs) (Table S4.1). Around half of the CNGs and CNLs were present in a maximum of two samples, with most of them being sample-specific, whereas the other half of the variants were distributed across a range of individuals (Figure 2a). At the chromosome level, the number of CNLs was consistently higher than the number of CNGs, except for chromosomes 5 and 7 (Figure 2b), whereas the genome-wide distribution of CNV variants was similar and covered the largest portion of the chromosomes (Figure 2d). Of the identified CNVs, 384 CNGs and 673 CNLs overlap with a total of 2942 annotated genes of the olive tree genome (Tables S4.2 and S4.3).

Structural variant (SV) calling was performed for olive tree genes plus 5 kb upstream and downstream of genes since we were interested in detecting variation within or close to genes that might be responsible for observed phenotypic variation. We produced a unified catalog of SVs that were called by at least two out of four bioinformatics tools and merged them using SURVIVOR (see 'Materials and Methods' section; Table 1; Table S4.4). In total, we detected 913 SVs (≥50 nt) (851 deletions and 62 duplications). Deletions had a median size of 1434 bp (standard deviation: 2535.3), ranging from 51 to 21 113 bp. Duplications had a median size of 1916 bp (standard deviation: 3186.4), ranging from 270 to 13 474 bp (Figure 2c). The distribution of deletions was uniform across the 23 chromosomes, whereas each chromosome had at least one duplication variant, except chromosomes 5, 19, and 21, on which we did not detect any duplication events. Detected deletions and duplications lie inside 1080 genes; their functional annotation is provided in Table S4.5. It is interesting to note that the genome-wide variations identified in this study were more abundant than those identified in previous studies (Jiménez-Ruiz et al., 2020; Julca et al., 2020), probably due to the large number of genotypes used in the present study.

Population structure and genomic variation among the olive tree genotypes

In the present study, the olive tree genotypes were divided into four groups, namely, NORTH-A MED, NORTH-B MED, NORTH-EAST MED, and WILD, based on the high-quality SNPs. The classification into these four groups was supported by the phylogenetic tree, unsupervised population structure analysis, and principal component analysis (PCA) (Figure 3b–d). As expected, all the olive tree cultivars formed a distinct group, separate from the wild olive tree genotypes and other species (Figure 3b,c). Similarly, phylogenetic inference split the olive tree cultivars into four distinctive groups, where indigenous varieties from the Mediterranean basin resided in two main groups, while the

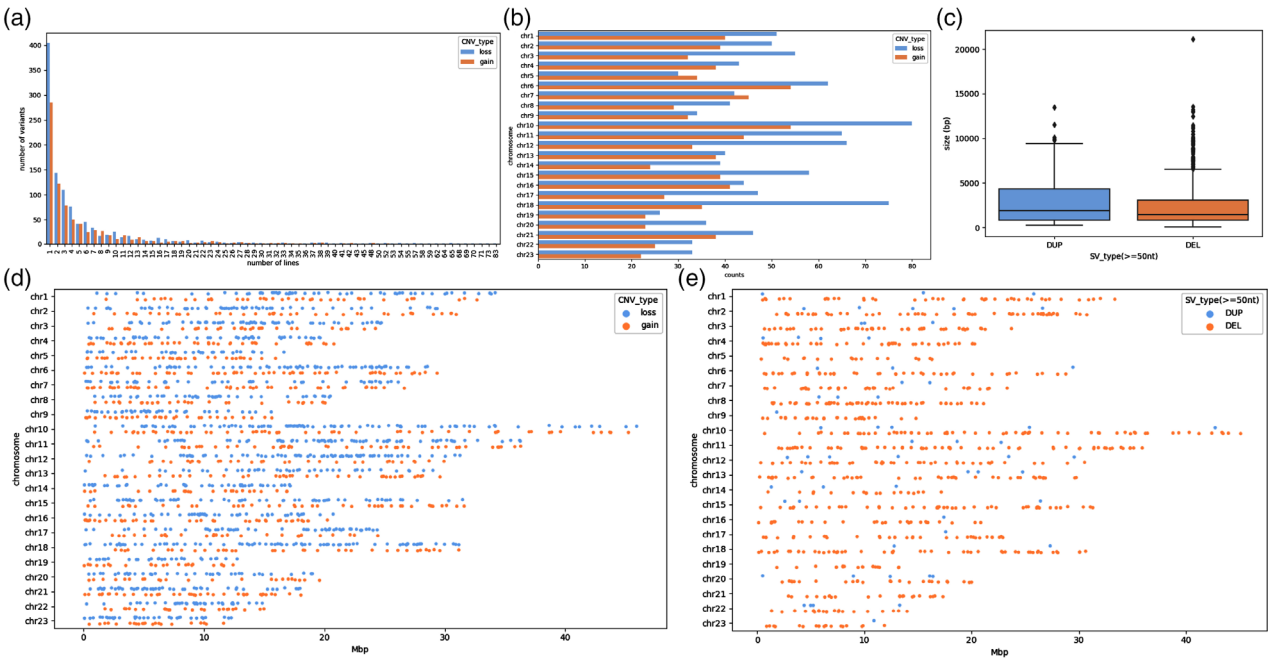

Figure 2. (a) Copy number variant (CNV) specificity. The number of individuals containing a variant was calculated for CNVs.
(b) Distribution of the number of CNVs across the 23 chromosomes.
(c) Distribution of CNVs across the 23 chromosomes.
(d) Size distribution of large deletions and duplications detected in genic regions.
(e) Distribution of large deletions and duplications (>50 bp) across the 23 chromosomes.

traditional cultivars originated from Iran, Syria, Greece, and Israel resided in the other.

Comparing the different olive tree groups, the abundance of SNPs was higher and the abundance of CNVs was lower in wild genotypes than in cultivars (Table S4). Moreover, a higher number of variations (SNPs and InDels) was observed in wild genotypes than in cultivars (Tables S5 and S7.1–S7.5). We also inferred the population structure using ADMIXTURE software and pre-defined the number of genetic clusters *K* from 2 to 12. The lowest cross-validation error was obtained when the value of *K* was 4. This analysis revealed the following four clusters: Cluster 1 (33 genotypes highlighted in black, primarily NORTH-A MED), Cluster 2 (14 genotypes highlighted in orange, primarily NORTH-EAST MED), Cluster 3 (11 genotypes highlighted in blue, primarily WILD) (Figure 3d), and Cluster 4 (29 genotypes highlighted in green, primarily NORTH-B MED), as well as the outgroup (two genotypes highlighted in purple). Notably, each cluster included some individuals with admixed ancestry, thus implying potential introgression. For example, in Cluster 1, 15 of the 33 individuals possessed a background from Cluster 4. The constructed unrooted neighbor-joining (NJ) tree and PCA of the 89 olive tree genotypes revealed a similar pattern with admixture clustering, demonstrating genetic discrimination for six clusters (Figure 3c). According to our ADMIXTURE analysis, NORTH-A MED, WILD, and NORTH-B MED populations were relatively well separated from each other on the first two PCA axes. On the contrary, the NORTH-EAST MED population was the only one that included samples that originated from the Southeast Mediterranean, occupying the quadrants on the right of the plot, partially overlapping with samples from groups NORTH-A MED and NORTH-B MED. Exploration of the third and fourth Principal Components (PCs) provided further insight as the two outgroup samples (OUT) were clearly distinguished from all the other individuals (Figure S1). The results of the NJ dendrogram (Figure 3c) were in agreement with the PCA results, although the group separation was less well defined, especially for the NORTH-B MED cluster. Interestingly, regarding the Greek gene pool, a North to South separation was observed, with northern Greek cultivars clustering to NORTH-A MED, while the rest of them formed the NORTH-B MED group. As depicted by the PCA, northern Greek cultivars formed a dense cluster within the NORTH-A MED population and showed low genetic differentiation among them (Figure 3b), possibly indicating their origin from a restricted gene pool.

As a further step, the phylogeny and migration history of cultivated and wild olive tree groups were investigated using TreeMix. Using *O. europaea* subsp. *cuspidata* as the outgroup, wild (WILD) genotypes were placed in one clade, and the three cultivar groups (NORTH-EAST MED, NORTH-A MED, and NORTH-B MED) were placed in a second clade

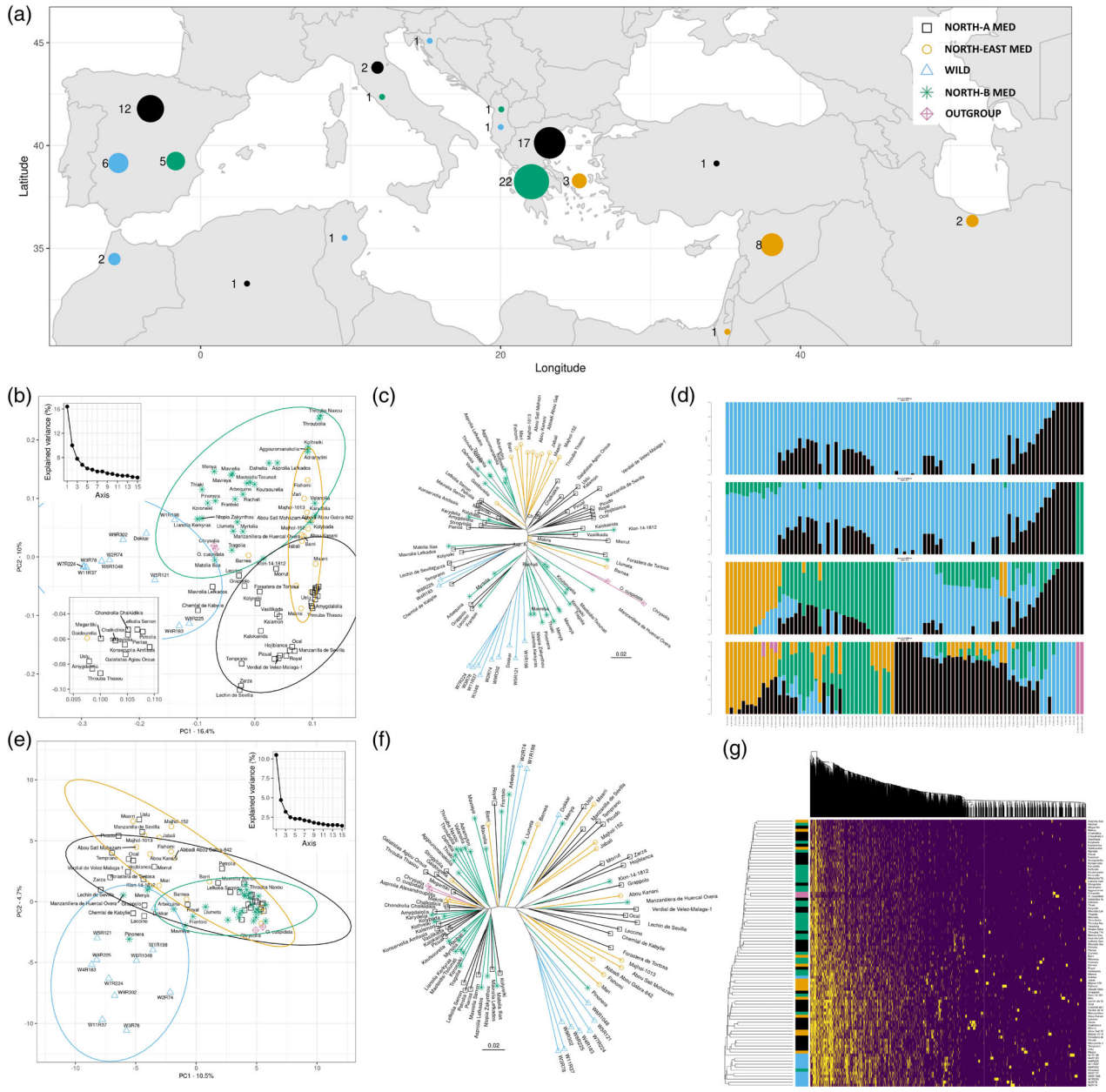

Figure 3. Population structure of *Olea europaea* genotypes in this study.
(a) Geographic origin of 89 *O. europaea* genotypes.
(b) First two principal component analysis (PCA) axes for the 89 *O. europaea* genotypes based on single nucleotide polymorphisms (SNPs).
(c) Neighbor-joining (NJ) dendrogram for the 89 *O. europaea* genotypes based on SNPs.
(d) Structure analysis. The results are shown for $K = 2$, $K = 3$, $K = 4$, and $K = 5$.
(e) First two PCA axes for the 89 *O. europaea* genotypes based on structural variant (SV) data.
(f) NJ dendrogram for the 89 *O. europaea* genotypes based on SV data.
(g) Cluster map for the 89 *O. europaea* genotypes based on SV data.

(Figure 4a). Additionally, we detected two gene flow events among the five groups: from WILD to NORTH-A MED and from NORTH-EAST MED to the outgroup (Figure 4a). Previous studies showed that all olive tree cultivars share a common nuclear ancestry from the East Mediterranean (Besnard et al., 2013; Julca et al., 2020), while most of them share a common eastern plastid lineage (Besnard et al., 2011, 2013; Julca et al., 2020). Based on these data, the first gene flow event potentially demonstrates that NORTH-A MED cultivars originated in the West and Central Mediterranean after introgressions between selected cultivated material from the East and wild genotypes, thus

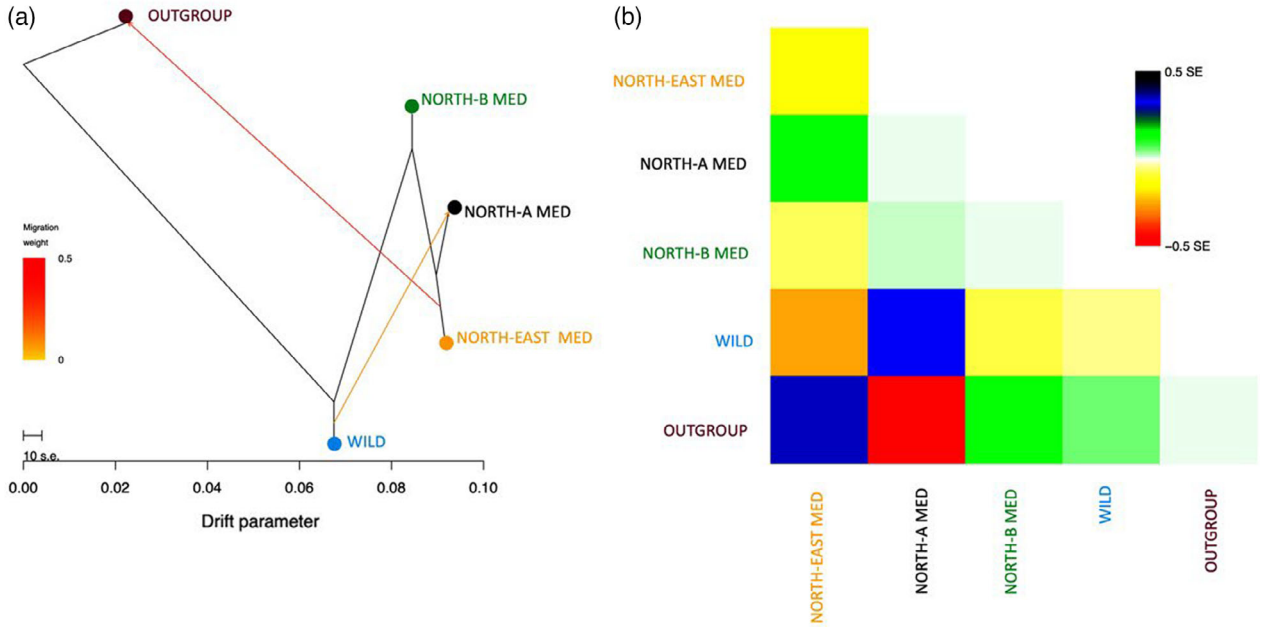

Figure 4. (a) Maximum likelihood tree produced with TreeMix.
(b) Residuals plot for the maximum likelihood tree with two migration edges.

confirming earlier domestication studies in olive tree (Gros-Balthazard et al., 2019; Julca et al., 2020). The second event could be attributed to gene exchange between the two subspecies during the historical spread of cultivars outside the Mediterranean basin, as reported by Besnard et al. (2018).

An interesting finding that emerged from this work is the fact that the well-known Greek cultivar 'Throubolia', which nowadays is grown mainly in Crete, was clustered with the monumental olive tree 'Throuba Naxos', showing an identical by descent value of 0.967, as it was determined by KING v.2.3.0 toolset (Manichaikul et al., 2010), within the NORTH-B MED group (Figure 3). Probably, the 'Throubolia' cultivar coincides with the monumental 'Throuba Naxos' olive tree originating from the Adisaros area of Naxos island, Greece (Kostelenos & Kiritsakis, 2017). However, it is very difficult to ascertain the precise age of an individual tree since olive trees are generally multi-stemmed, particularly in wild habitats (Baali-Cherif & Besnard, 2005). In our study, based on the radius and the circumference of the Naxos olive tree (Koniditsiotis, 2020), we estimated the tree's age to be about 3000 years (specifically, 2915 or 2918 years, depending on the equation used). Notably, its olives (known as 'thrubes') having ripened on the tree are edible without further processing, a trait that makes the 'Throubolia' olive fruit unique (Kostelenos & Kiritsakis, 2017). The existence of this monumental tree with its unique agronomic trait in group NORTH-B MED might possibly indicate that olive tree domestication in Greece dates back to at least 3000 years ago.

Overall, our results are in congruence with recent genomic (Julca et al., 2020) and transcriptomic (Gros-Balthazard et al., 2019) studies and support the hypothesis of a main domestication event in the eastern part of the Mediterranean basin, followed by dispersion towards the West accompanied by secondary introgression events with wild olive trees. Future studies involving a more representative sample of wild individuals throughout the Mediterranean could provide a higher-resolution overview of the extent of these events and possibly elucidate whether a single or several independent domestication events occurred in olive tree.

Genetic diversity and differentiation among olive tree groups

The total non-synonymous/synonymous mutation ratio (Ka/Ks) was 1.45 (1.44 for NORTH-A MED, 1.42 for NORTH-EAST MED, 1.43 for WILD, and 1.49 for NORTH-B MED) (Table 1; Table S5). These values were comparable to those observed in other fruit tree species such as mango (*Mangifera indica*) (1.52; Wang et al., 2020) and sweet cherry (1.78; Xanthopoulou et al., 2020), which are usually propagated by grafting or clonally. Nonetheless, this value is much higher than the values reported in peach (*Prunus persica*) (1.06; Li et al., 2019). The higher Ka value obtained in our olive tree dataset is probably the result of the artificial selection pressure imposed by breeding in combination with the vegetative propagation through grafting of domesticated cultivars (Xanthopoulou et al., 2020). At the level of the whole genome, the occurrence of higher non-

synonymous-to-synonymous substitution ratios in cultivated olive tree genotypes compared to wild genotypes suggests that the cultivated olive trees have accumulated a higher ratio of deleterious to non-deleterious mutations (Varshney et al., 2017).

The values of two summary statistics, nucleotide diversity (π) and theta (Watterson estimator, θw), were calculated to investigate genome-wide patterns of polymorphism as well as to estimate the genetic diversity for the different olive tree groups (Figure 1b). Among them, it was evident that the WILD group displayed higher genetic diversity than the cultivated olive tree groups (NORTH-A MED, NORTH-B MED, and NORTH-EAST MED). Meanwhile, olive tree cultivars from the WILD group exhibited the highest level of genetic diversity ($\pi = 7.08 \times 10^{-4}$ and $\theta w = 7.34 \times 10^{-4}$), followed by NORTH-A MED, NORTH-EAST MED, and NORTH-B MED (Figure 1b). Based on these results, comparisons of π values between the different groups showed consistent diversity between the cultivated gene pools. It is noteworthy that when comparing the diversity values (θw and π) of the wild (WILD) group with the ones of the cultivated groups (NORTH-A MED, NORTH-EAST MED, and NORTH-B MED), we noticed an opposite pattern with lower θw and π values (Figure 1b). The reduction in the θw and π values observed in the cultivated olive tree groups could be due to loss of segregation sites caused by genetic drift during the domestication process, while the heterogeneity across genome-wide loci was increased, possibly due to artificial hybridizations as a result of olive tree breeding.

Present data further indicated that the Tajima's *D* values for the WILD group (0.076) were much lower than those for the cultivated groups (1.335, 0.923, and 1.217 for NORTH-A MED, NORTH-EAST MED, and NORTH-B MED, respectively), which could be attributed to the fixation of alleles driven by genetic drift. This difference between wild and cultivated genotypes may be the result of the propagation method, the admixture, or the population reduction. On the contrary, negative values of Tajima's *D* at local regions in the genome have also been reported for domesticated groups of cultivated olive trees, supporting the hypothesis of strong selection for specific alleles during the domestication process (Jiménez-Ruiz et al., 2020). Therefore, our results uncovered interesting aspects of possible recurrent genetic admixture events between wild populations and cultivated germplasm across the Mediterranean basin that merit further investigation.

As a further step, the Wright fixation index (F_{ST}) was calculated to estimate the genetic differentiation among the olive tree populations. Initially, the differentiation level between the wild olive trees (WILD) and the outgroup (OUT) was assessed, as this may provide some valuable insights into the breeding history and origin of diverse cultivars. Our F_{ST} and TreeMix results were congruent (Figure 4a) and demonstrated that the WILD group of genotypes exhibited a significant level of differentiation from the OUT group ($F_{ST} = 0.424$) (Figure 1b). The differentiation between the wild and cultivated groups presented its lowest values between WILD and NORTH-B MED ($F_{ST} = 0.167$), followed by NORTH-A MED ($F_{ST} = 0.201$) and NORTH-EAST MED ($F_{ST} = 0.241$). The overall F_{ST} value between wild relatives (OUT) and NORTH-EAST MED cultivars was slightly higher (0.386) than the F_{ST} value between OUT and NORTH-A MED (0.384) or NORTH-B MED cultivars (0.374) (Figure 1b). These data suggest that the divergence between WILD genotypes in our study and NORTH-EAST MED (0.241) cultivars was lower than previously reported by Julca et al. (2020) (0.410), possibly due to the higher number of cultivars and wild genotypes examined in the present study.

In addition to population differentiation due to genetic structure, we also focused on the linkage disequilibrium (LD) of the olive tree cultivars, which is known to be important for the genetic diversity analysis of olive tree germplasm. Analysis of LD for the four olive tree clusters indicated that the LD decay rate was high for all olive tree genotypes, with 18.5 kb (herein measured as the distance at which the average pairwise SNP correlation coefficient r^2 dropped to half of its maximum value) for the entire population (Figure 1c). Consistent with the reduction in genetic diversity detected in the NORTH-EAST MED population, this group had a much slower decay rate (41 kb) than the WILD group (24 kb). More rapid LD decay was observed in the cultivated genotypes originating from the northern Mediterranean region compared to the eastern Mediterranean genotypes. This could be due to the fact that the NORTH-EAST MED population includes geographically distinct accessions, i.e., originating from Greece, Spain, Italy, Croatia, Albania, Turkey, Morocco, and Tunisia. Previous data also showed that LD decreased with physical distance among SNPs in 57 olive tree cultivars (Zhu et al., 2019). The rapid LD decay in the cultivated olive trees may be useful for identifying the associated QTL intervals that resulted in the identification of less spurious candidate genes in GWAS experiments.

Diversity in olive tree genotypes based on SVs

Similar to SNPs, SVs may serve as a measure of genetic diversity, useful to investigate the evolutionary relationships within olive tree groups (Alonge et al., 2020). To analyze these relationships, in the current study three approaches were used: (i) clustering, (ii) NJ, and (iii) PCA. All approaches were based on the combination of common and unique SVs present in the 89 genotypes (Figure 3e–g). The PCA plot based on the combined dataset for all types of SVs indicated the presence of a well-separated group mostly containing wild individuals, situated in the lower left quadrant. By contrast, the rest of the

groups were not well separated. Notably, samples of Greek origin were found exclusively in the right section of the plot forming a tight cluster. Interestingly, the two outgroup genotypes were also present near this cluster. The NJ dendrogram corroborated the PCA results. Greek olive trees formed separate clusters from the rest of the Mediterranean cultivars. The single Croatian wild sample and one Spanish wild genotype clustered closer to the Mediterranean cultivars than the rest of the wild individuals. Taken together, both the SVs and the CNVs showed consistent phylogenies (Figure 3f), suggesting the validity of these data.

Signatures of selection in olive tree genomes

Using the proportion of variance explained and projecting individuals on the principal components as a score plot, we estimated that the optimal number of PCs from the SNP matrix was three. Plots of loadings for the selected axes displayed a homogeneous distribution in all genomic regions, indicating no bias as a result of LD. Figure 5 depicts the Manhattan plots illustrating the SNPs identified as being under selection pressure on all olive tree chromosomes according to a PCadapt test. Nine regions were identified to be under selection by the applied statistics and were defined as candidate selection sweep regions. These regions were found on chromosomes 3, 4, 10, 12, 13, and 17 (Table S8).

A high number of the selected SNPs fall into genes with known functions. For example, the SNP at 85.1 Mb on chromosome 3 is located within the promoter of the gene 'Oeu010785.1', which is related to root and leaf hydraulic conductance, as the expression of PIP aquaporins (AQPs) is enhanced in olive tree dwarf genotypes (Lovisolo et al., 2007). In addition, Faize et al. (2020) suggested AQPs

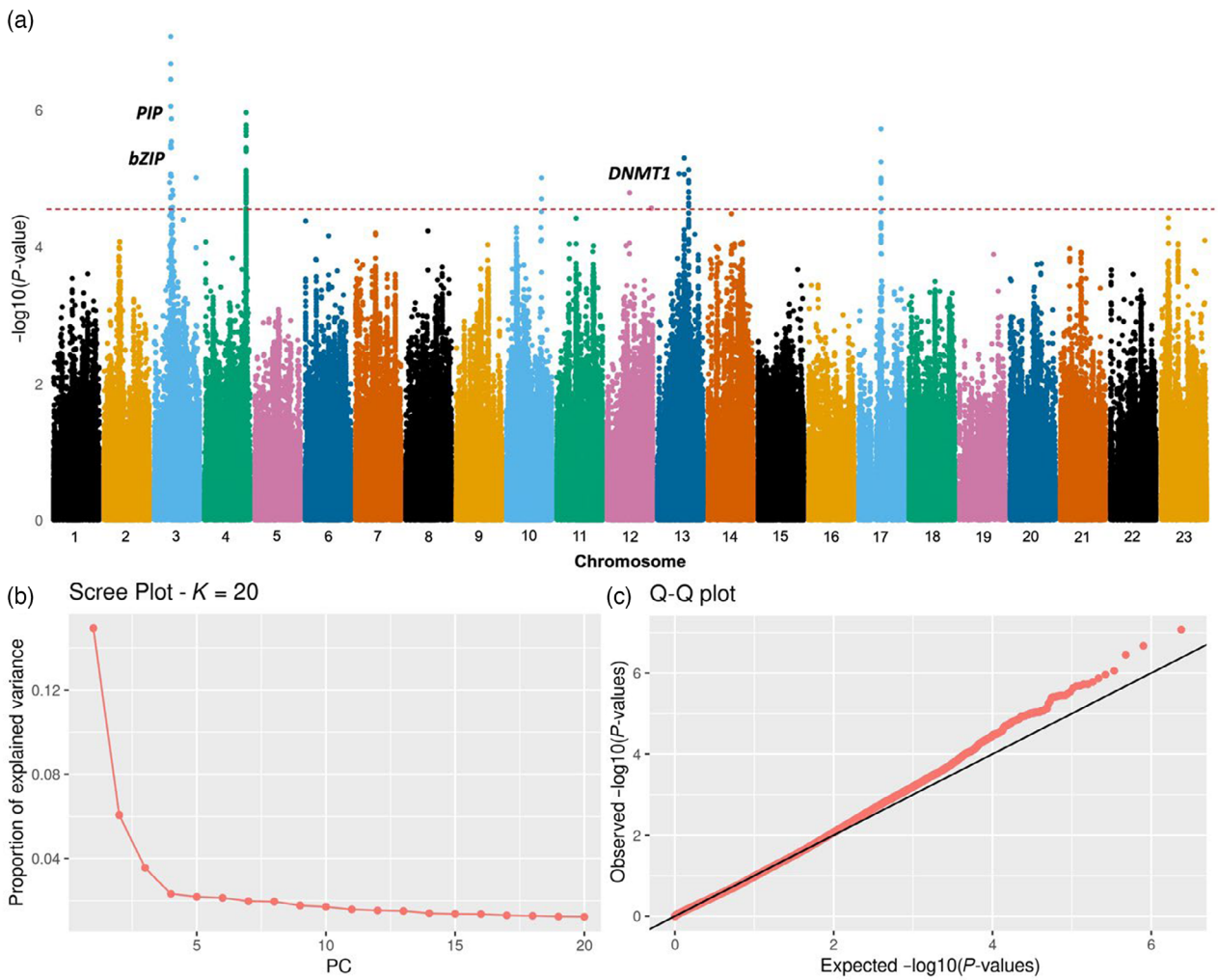

Figure 5. Signatures of selection in olive tree genomes using *pcadapt*.
(a) Manhattan plot of adjusted *P*-values (*q*-values) showing genomic regions comprising single nucleotide polymorphisms (SNPs) under selection pressure.
(b) Scree plot displaying the proportion of variance explained for the first 20 principal component analysis axes containing neutral and adaptive SNPs.
(c) Quantile-quantile plot of observed versus expected transformed *P*-values.

play a role in the adaptation of olive trees to diverse environmental conditions in order to support the genetic improvement of domesticated olive trees. Similarly, SNPs belonging to the sweep region detected in this study on chromosome 3 are located in the coding region of 'Oeu042954.1' (encoding an OebZIP transcription factor), which is related to olive fruit development and lipid synthesis (Rong et al., 2020). Furthermore, the gene encoding DNA methyltransferase 1 (DNMT1), which is located on chromosome 12, was identified as a locus under selection. DNMT1 has long been considered as one of the major maintenance methyltransferases (Corominas-Faja et al., 2018) and is related to the phenol content of extra virgin olive oil (EVOO). Moreover, on chromosome 13, a region under selection includes the gene *ASA* (Oeu019857.1), which is involved in tryptophan (Trp) biosynthesis (Sato & Matsui, 2011) (Table S8).

Variation in genes involved in flowering time, self-compatibility, and fruit weight

Another interesting result of the present study was the selection of 12 candidate genes (from previous QTL studies in olive tree genotypes) related to flowering time, fruit weight, and self-compatibility (Kaya et al., 2019; Mariotti et al., 2020; Moret et al., 2023) to identify annotated SNP variants with a high impact on the functions of those genes. All SNPs identified in the 12 candidate genes are reported in Table S9. This analysis identified high-impact mutations in flowering time genes, including a mutation resulting in a stop codon gain (G/T; Oeu055051.1) in the *AP2*-like gene. It was proposed that the downregulation of *AP2*-like target genes by *miR172* during early seedling development in wild-type Arabidopsis relieves floral repression by these genes, resulting in the promotion of flowering (Aukerman & Sakai, 2003). In addition to this, we found four mutations resulting in stop codons in *STERILE APETALA* (*SAP*) genes (Oeu021293.1); one of them was detected in the flowering locus T (*FT*) gene. The present observations are not without precedent, as previous work indicated that overexpression of the *FT* gene is linked to early flowering in olive tree (Haberman et al., 2017). It was previously shown that a recessive mutation in the *SAP* gene can provoke severe alterations in inflorescence and flower and ovule development in Arabidopsis (Byzova et al., 1999). Apart from putative flowering-related genes, we identified several genes associated with self-compatibility, including the self-compatibility-associated transmembrane and coiled-coil domain-containing protein 4 gene (Oeu037736; G/A). Mariotti et al. (2020) mapped the diallelic self-incompatibility locus (interval 5.4 cM on linkage group 18) in the olive tree genome and reported that this region has a length of approximately 300 kb. To this end, we noted high-impact mutations in genes associated with fruit weight. According to Kaya et al. (2019), two candidate genes on chromosome 15 are linked to fruit weight (Oeu060693.1 and Oeu060694.1). Congruent with the aforementioned study, we identified three and 11 high-impact SNPs in Oeu060693.1 and Oeu060694.1 (both encode polynucleotidyl transferase), respectively (Table S9), thus suggesting a role for these genes as effectors of olive tree growth. Recently, Moret et al. (2023) have also identified several genes that are associated with olive fruit weight through GWAS and RNA sequencing, suggesting this trait is polygenic. Their study highlighted two genes that appear to be involved in the early stages of cell division, four genes that are responsible for regulating fruit development, including one gene that encodes a long non-coding RNA, and a gene that may be involved in fruit cell expansion. Among these seven genes, we identified modifier SNPs in the promoter regions of Oeu007256.2 (Armadillo repeat-containing protein 6) and Oeu018720.1 (fructose-bisphosphate aldolase, class I) (Table S9). No other high-impact SNP was identified within the seven genes among the 89 olive tree genotypes.

Genome-wide associations with agronomic and morphological traits

Despite the wide application of GWAS in cereals and other annual crops, only a few GWAS have been conducted for perennial fruit trees (Zhang et al., 2021). Another major objective in this study was, therefore, the genome-wide association mapping of agro-morphological traits among a diverse collection of Greek-oriented olive tree cultivars. To achieve such challenge, a total of 23 morphological traits were phenotyped, including phenotypic categories like flower number, fruit weight, and fruit shape and two traits with agronomic interest (olive oil content and ripening time) (Table S2; Figure 6a). Spearman correlation analysis of the 23 morphological traits (Figure 6a,b) showed a positive correlation between endocarp and fruit traits such as weight ($r = 0.75$, $P = 3.61e{-}10$), shape ($r = 0.47$, $P = 5.47e{-}4$), and symmetry ($r = 0.41$, $P = 3.38e{-}3$). Interestingly, a positive correlation was found for leaf length with fruit and endocarp weight (Figure 6a,b).

A GWAS was performed aiming at identifying associations between genetic variation and agronomic and morphological traits. The association panel was composed of 50, 48, and 41 out of 77 cultivars for morphological traits, olive oil content, and ripening time, respectively (Table S2). GWAS was performed with the SNP variants filtered for minor allele frequency (MAF) (≥0.1), minor allele number (≥5), and missing rate (≤0.5), leaving a total of 428 320 SNPs. In order to avoid the detection of false positive alleles, the Bonferroni threshold was used to identify the significant peaks. For every trait, Manhattan plots were generated to visualize SNPs that were significantly associated with morphological and agronomic traits across the *O. europaea* genome.

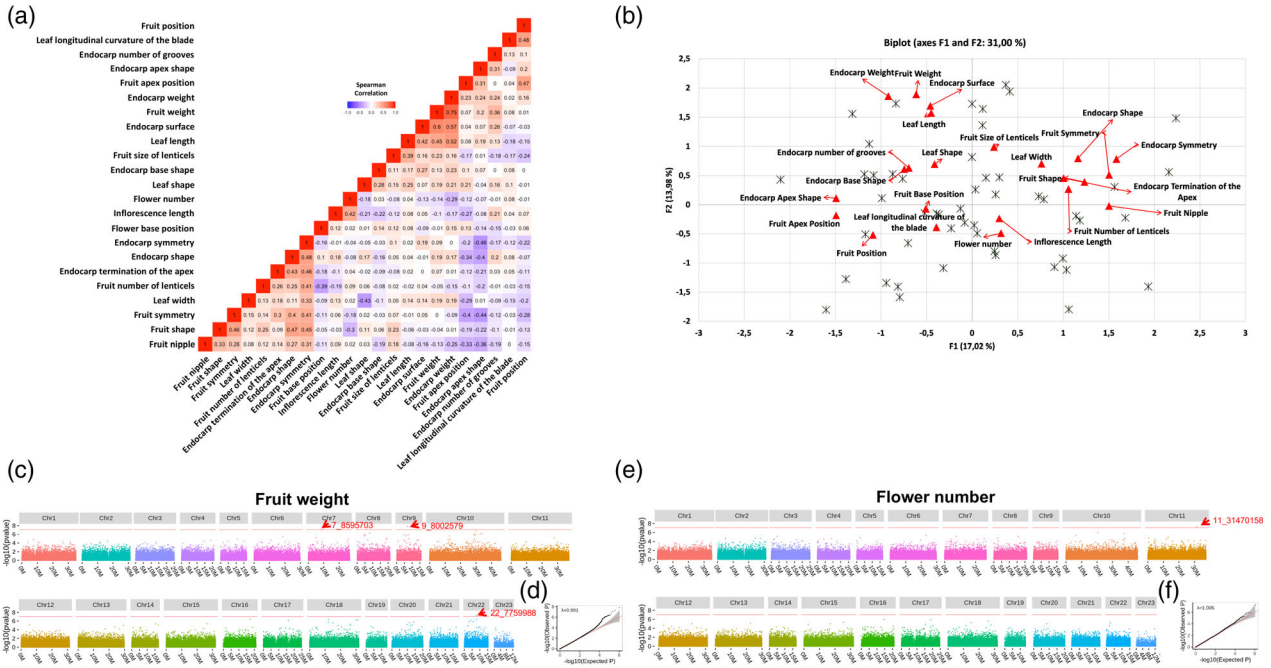

Figure 6. (a) Spearman correlation analysis showing correlation among the 23 morphological traits.
(b) Principal component analysis (PCA) biplot based on the phenotypic traits of the 50 olive tree genotypes regarding the first two principal components (F1 and F2). Biplot PCAs were performed for genotypes and traits using XLSTAT software (version 2014.1).
(c–f) Manhattan and quantile-quantile (QQ) plot of genome-wide association scan for (c, d) fruit weight and (e, f) flower number in 50 olive tree cultivars. Log transformed *P*-values are plotted against the physical position on the respective chromosomes for 428 320 single nucleotide polymorphisms with minor allele frequency > 0.1. The horizontal red line indicates the Bonferroni threshold.

Across all the analyzed traits (Table S2), significant associations were detected for 'fruit weight', 'flower number', 'apex position', 'base position', 'leaf width', 'leaf length', 'size of lenticels', and 'fruit shape' (Figure 6; Figures S2 and S3). The positions of all significant SNPs were used to further investigate these loci for candidate genes potentially associated with the phenotypes. In particular, the locations of the significant SNPs were determined by using the *O. europaea* var. *sylvestris* genome annotation (Unver et al., 2017) and regions were scanned for annotated genes within a 100-kb interval. A list of all annotated genes tagged by the significant SNPs for every trait is presented in Table S10. In addition to the SNPs that tagged previously annotated genes, many loci consisted of a large number of genes with unknown function.

Flower number and fruit weight are commercially very important traits in olive tree cultivation, since they play a crucial role in fruit load, fruit size, and yield (Moret et al., 2023; Rosati et al., 2012). For both flower number and fruit weight, we have identified novel, significant loci (Figure 6). Specifically for fruit weight, the significant peak found in chromosome 22 tagged a gene (Oeu005088.4) annotated as COP1-interacting protein 7-like (Table S10). Although the active role of the *COP1* gene in fruit weight is still unknown, a previous study found that COP1 negatively regulates hypocotyl elongation and anthocyanin biosynthesis in apple (*Malus* × *domestica*) (Kang et al., 2022). By further examining the functional annotation of the other SNPs identified in this gene, we have found non-synonymous SNPs that are annotated as 'missense_variant; splice_acceptor_variant; splice_donor_variant; splice_region_variant; start_lost'. These SNPs suggest that this gene might play a role in olive fruit weight. Regarding the other three loci on chromosomes 7, 9, and 22, although the SNPs fall into intergenic regions, we have found five candidate genes within a 100-kb interval, encoding a polyol transporter (Oeu033162.1), dehydration-responsive element-binding protein 2F (Oeu033165.1), casein kinase 1-like (Oeu005089.2), and signal recognition particle 14 kDa protein-like (Oeu017427.1), as well as Oeu032434.1, a gene with unknown function (Table S10). However, the absence of non-synonymous SNPs within these five genes makes them weaker candidates for being associated with fruit weight than Oeu005088.4. Nevertheless, these five candidate genes that were identified in the present study in combination with the seven previously identified genes by Moret et al. (2023) that are potentially associated with olive fruit weight can significantly contribute to the breeding of large-sized olive fruit cultivars.

Flower number represents a trait with high significance for yield and consequently olive fruit size (Rosati et al., 2012). Association analysis of the present dataset

identified a locus on chromosome 11 (Figure 6). The most significant SNP is located in an exon of the alpha-carbonic anhydrase 1 (*α-CA1*) gene (Oeu047863.1) and remarkably it causes a non-synonymous mutation (Table S10), suggesting that this mutation makes *α-CA1* a very strong candidate gene for flower number. Although the extent to which *α-CA1* influences olive tree flowering, directly or indirectly, is unclear, this could be the result of the complexity surrounding multiple α-CA1 isoforms found in photosynthetic tissue, their different abundances in various organelles, and their potential roles in multiple metabolic pathways (DiMario et al., 2018; Rudenko et al., 2021). Other genes that are found within the 100-kb interval are annotated as 'myb-related protein 1-like' (Oeu047864.2), 'replication protein A' (Oeu047859.1), 'ABC transporter B family member 29' (Oeu047858.1), and 'Oeu047861.1', a gene with unknown function (Table S10). Hence, this study provides insight into a yet unknown mechanism underlying the flowering process that will constitute a significant contribution to our knowledge of olive tree physiology.

Apart from the two important traits of 'fruit weight' and 'flower number', a GWAS was conducted for 22 more traits (Figures S2 and S3). The association analyses identified significant loci for six ('base position', 'fruit size lenticels', 'leaf length', 'apex position', 'fruit shape', and 'leaf width') out of those 22 traits. As a further step, the location and functional annotation of significant SNPs were studied to reveal genes of interest within the 100-kb interval. Interestingly, several SNPs are located within coding or promoter regions (Table S10). These candidate genes need to be further investigated in order to be exploited for future olive tree breeding efforts.

CONCLUSION

Herein, we presented the most complete genomic variation map to date for olive trees originating from the entire Mediterranean basin. WGRS analysis focused on genomic regions associated with favorable variation, such as CNVs, deletions, substitutions, and duplications, providing the first comprehensive catalog of molecular variation in *O. europaea*. We discovered numerous allelic variants with a high impact on flowering time and fruit weight genes as well as genes linked to key agronomic traits that can be employed for functional analyses and future breeding efforts. By using the monumental tree 'Throuba Naxos' we shed light on the potential origin or routes of olive tree domestication. Taking into account the unique agronomic traits, the key location (middle of Aegean Sea), and the clustering of this monumental olive tree in the NORTH-B MED group, we could possibly conclude that olive tree domestication in Greece dates back to at least 3000 years ago. Several candidate genes were uncovered by a selective sweep scan to be under selection pressure on all olive tree chromosomes. These genes are known to be associated with key agronomic traits such as adaptation to diverse environmental conditions, fruit development, lipid synthesis, and the phenol content of EVOO. To further exploit the genomic and phenotypic resources obtained from the current work, genome-wide association analyses were performed for 23 morphological and two agronomic traits. Significant associations were detected for eight traits that provide valuable candidates for olive tree breeding for enhanced olive oil quality and fruit yield. Our study provides insights into the complex olive tree domestication and biology and identifies genes controlling important traits in olive tree, which extends the genetic resources and provides a basis for molecular breeding in this high-value tree species.

EXPERIMENTAL PROCEDURES

Plant material and site description

Herein, we studied 504 olive trees belonging to 41 *O. europaea* subsp. *europaea* cultivars from Greece (40) and Spain (one), one monumental olive tree, and two ornamental genotypes of *O. europaea* subsp. *cuspidata*. All 41 cultivars are maintained in the NOGB located at the Chrisopigi Monastery area near the Institute of Olive Tree, Subtropical Crops & Viticulture (I.O.S.V.), Hellenic Agricultural Organization 'ELGO – Dimitra' (Chania, Southern Greece) (Table S1), including a presumably 'ancient' monumental tree from Naxos Island and the two genotypes of *O. europaea* subsp. *cuspidata*. We also included a group of 45 out of 52 olive tree genotypes reported in a previous work (Jiménez-Ruiz et al., 2020) that consists of 35 olive tree cultivars belonging to an established core collection (Belaj et al., 2012), as well as six other economically important cultivars and 10 wild genotypes (Table S1). In total, we included 89 unique genotypes from both datasets.

For GWAS, observations were collected over a period of 20 years (1997–2017). Missing data were associated with years of no fruit production. Nevertheless, we ensured that we obtained phenotypic measurements for three consecutive years of fruit production for all cultivars. Depending on availability, two to 12 trees per cultivar were sampled. All phenotyping parameters were averaged over multi-year measurements to minimize unaccountable variance. Mean air temperature in this area was 18°C, mean relative humidity was 64%, and annual rainfall was 660 mm (H.A.O. Meteorological Station, Chania, Greece). The following physicochemical properties of the upper soil layer (0–30 cm) were recorded: clay, 244–284 g kg^{-1}; silt, 220–260 g kg^{-1}; sand, 476–510 g kg^{-1}; pH 6.63–7.01; conductivity, 0.14–0.22 dS m^{-1}; total $CaCO_3$, 0.12–0.18%; soil organic matter, 0.40–1.34%.

Age estimation of the Naxos ancient olive tree

The age of the Naxos ancient olive tree was estimated using the following two equations:

$$y = 5.2983x + 54.431, \quad (1)$$

$$y = 26.87 + 99.7 \times P, \quad (2)$$

where x is the radius in centimeters at 1 m of height and P is the circumference in meters (Koniditsiotis, 2020; Pannelli et al., 2010). Considering that the radius at 1 m height and the circumference of the Naxos olive tree are 540 cm and 29 m, respectively, the

estimated age is 2915.5 and 2918.17 years according to the first and the second equation, respectively.

Morphological description of the olive tree cultivars

The methodology of the RESGEN project as presented in the World Catalogue of Olive Varieties (Barranco et al., 2000) was employed for the morphological description of olive tissues, namely leaves, inflorescences, fruits, and endocarps, during the period of 1997–2001 (5 years). For each cultivar and year, 40 mature leaves were collected from the middle section of eight to 10 1-year-old shoots. Inflorescence traits were observed from a sample of 40 inflorescences at the white bud stage, taken from the middle section of eight to 10 fruiting shoots (previous year growth). A sample of 40 fruits was collected from the middle section of fruiting shoots when color change had been completed and then used to evaluate endocarp traits. Leaves, inflorescences, fruits, and endocarp samples were taken from typical shoots on the south-facing side of the trees. The following 32 traits were evaluated by a group of three trained experts: four tree traits (vigor, growth habit, canopy density, and length of internodes), four leaf traits (shape, length, width, and longitudinal curvature of the blade), two inflorescence traits (length and number of flowers), 11 fruit traits (weight, shape in position A, symmetry in position A, position of maximum transverse diameter in position B, apex in position A, base in position A, nipple, presence of lenticels, size of lenticels, location of start of color change, and color at full maturity), and 11 endocarp traits (weight, shape, symmetry in position A, symmetry in position B, position of maximum transverse diameter in position B, apex in position A, base in position A, surface in position B, number of grooves, distribution of the grooves, and termination of the apex in position A) (Table S2). A preliminary description of endocarp traits for 42 olive tree genotypes was presented in Koubouris et al. (2019).

Determination of ripening time

For each cultivar and year, 40 fruits were collected from the middle section of fruiting shoots every week from the middle of September until the end of December during the period of 2002–2004 (3 years). Ripening time was determined according to the methodology of the RESGEN project as presented in the World Catalogue of Olive Varieties (Barranco et al., 2000) based on the date on which the first fruit change color occurred (1–10% of the total number of fruits) from green to purple in the epidermis, corresponding to phenological stage 81 (Sanz-Cortés et al., 2002).

Fruit oil content

Olive oil content determination was performed by using two subsamples of each cultivar for 4 years (2001–2004). Each subsample consisting of 50 g of fruit flesh and stone was dried to constant weight at 80°C, and water content was calculated. Oil content in crushed fruit drupes (dry weight) was determined gravimetrically by means of Soxhlet extraction according to Harwood and Moody (1989).

DNA isolation and whole genome resequencing

For genome resequencing, high-quality DNA was extracted from young leaves from a single tree of each accession using the DNeasy Plant Pro Kit (Qiagen, Inc., Valencia, CA, USA) according to the manufacturer's instructions. After dilution to 100 ng μl^{-1}, the 44 genomic DNA samples were used to generate 44 Illumina libraries with a mean insertion size of 500 bp. The libraries were sequenced using 150-bp paired-end reads using an Illumina NovaSeq 6000 system (Illumina, San Diego, CA, USA).

Variation calling and annotation

The quality of paired-end Illumina reads was evaluated using FastQC (v 0.11.9) and MultiQC (v 1.9). Illumina paired-end reads were pre-processed using fastp (v 0.20), where sequencing adapters and low-quality ends (<Q20) were trimmed. Trimmed reads meeting the filtering criteria (average quality ≥ Q20, minimum length ≥ 18 bp) were aligned to the reference genome of *O. europaea* var. *sylvestris* (Unver et al., 2017) using the Burrows–Wheeler alignment tool 'bwa mem' (v 0.7.17). PCR duplicates were marked using Picard (v 2.22.8). Mapping quality, sequencing depth, and coverage uniformity of read alignments are reported (Table S3). Multi-mapped reads were removed using samtools (v1.11) (Li et al., 2009). Variants were called using the Genome Analysis Toolkit (GATK). Specifically, the tools HaplotypeCaller, CombineGVCFs, GenotypeGVCFs, Select Variants, and Variant Filtration were employed and small variants identified were classified as SNPs only or InDels only. High-quality SNP variants were obtained using the following filters: --filter-name 'QD2' --filter-expression 'QD < 2.0' --filter-name 'QUAL30' --filter-expression 'QUAL < 30.0' --filter-name 'SOR3' --filter-expression 'SOR > 3.0' --filter-name 'FS60' --filter-expression 'FS > 60.0' --filter-name 'MQ40' --filter-expression 'MQ < 40.0' --filter-name 'MQRankSum-12.5' --filter-expression 'MQRankSum < −12.5' --filter-name 'ReadPosRankSum-8' --filter-expression 'ReadPosRankSum < −8.0' --filter-name 'ExcessHet-54.69' --filter-expression 'ExcessHet > 54.69' --genotype-filter-name 'GQ-30' --genotype-filter-expression 'GQ < 30' --genotype-filter-name 'DP-4and3000' --genotype-filter-expression 'DP < 3000 || DP ≥ 4'. Similarly, InDel variants were obtained using the following filters: --filter-name 'QD2' --filter-expression 'QD < 2.0' --filter-name 'QUAL30' --filter-expression 'QUAL < 30.0' --filter-name 'FS200' --filter-expression 'FS > 200.0' --filter-name 'ReadPosRankSum-20' --filter-expression 'ReadPosRankSum < −20.0' --filter-name 'ExcessHet-54.69' --filter-expression 'ExcessHet > 54.69' --genotype-filter-name 'GQ-30' --genotype-filter-expression 'GQ < 30' --genotype-filter-name 'DP-4and3000' --genotype-filter-expression 'DP < 3000 || DP > 4'. Variants belonging to individual samples were extracted using bcftools (v1.11), while variant annotation was conducted using SnpEff (v 4.3) (Cingolani et al., 2012).

Sequence and variant quality checking

For validation purposes, 16 SNPs were sequenced using Sanger sequencing technology in 16 *O. europaea* subsp. *europaea* cultivars. The 16 SNPs belong to the region of the iridoid synthase (*OeISY*) gene, which was amplified by PCR using gene-specific primers that we designed. For better accuracy, three sets of primers were designed for three overlapping subregions of *OeISY*: primer pair A, 5′-GGCATGTAGTTCTTGTTGATTC-3′ and 5′-GTGTGTCGTGTGATTCTGGTG-3′; primer pair B, 5′-TACCAAATGCTCCAAATCTCCA-3′ and 5′-CCCACCATCTCCAAGTCAAA-3′; primer pair C, 5′-AAGGCTTCGTGGAATTGTTTC-3′ and 5′-TTGCCCATTGTGAGCTGATTAG-3′. Both forward and reverse strands were sequenced. Sequencing chromatograms were analyzed using SnapGene® software, the three overlapping subregions were assembled by CAP3 using default parameters, multisequence alignment was performed using MUSCLE with default parameters, and reference genome alignment was performed with the SnapGene® integrated alignment tool.

Genetic variation in genes related to traits of interest

Candidate genes involved in the regulation of flowering time, self-compatibility, and fruit weight were selected. Genetic variability of the candidate genes was explored across the different genotypes and the potential effect of the genetic changes was studied using snpEff v.4.3 by annotating each SNP based on their predicted effect on the candidate genes.

SV calling and filtering

CNV at the genome level was analyzed using Control-FREEC v11.6 (Boeva et al., 2012) with control parameters. The output was filtered by applying a cut-off of $P < 0.05$ (Kolmogorov–Smirnov test). Consecutive variants of the same type were merged if they were less than 1 kb apart. Only CNVs with a maximum of 30 copies were kept. CNV variants were classified as CNG and CNL following Control-FREEC annotations. Since we were interested in detecting structural variation in genes that could explain an observed phenotype, we searched for large deletions and insertions (>50 bp) in annotated olive tree genes, including the 5-kb region upstream and downstream of each annotated gene. For SV detection in the specific regions, we used DELLY (Rausch et al., 2012; v0.8.1), LUMPY (Layer et al., 2014; v0.2.13), Manta (Chen et al., 2016; v1.6.0), and Platypus (Rimmer et al., 2014; v0.8.1). All software programs were run with default parameters, unless mentioned otherwise. In order to remove variants of low quality in LUMPY results, we filtered out sites that had a variant quality of less than 20 (QUAL < 20) and less than five supporting reads for the variant (SU < 5). Moreover, we filtered out sites with an MAF of less than 0.3 and sites where less than 5% of the individuals had a genotype assigned. We further reduced false positive variants with SURVIVOR (Jeffares et al., 2017; version 1.07), where we kept variants that were present in at least two software programs and merged those with a distance of 500 bp or less.

Genome scanning for selective sweep signals

The filtered VCF file was converted to bed format by PLINK 1.9 (Chang et al., 2015). The R package pcadapt (Privé et al., 2020) was utilized to detect outlier SNPs. The optimal choice of K was determined based on the scree plot and the projection of samples on the first three PCA axes (score plot). A false discovery rate procedure as implemented in the package q value (Storey et al., 2020) was used to indicate the cut-off for outlier detection for $a = 0.1$. The R commands for this analysis are available in Appendix S1.

Population genetics analyses

We used ADMIXTURE to analyze the admixture patterns for all olive tree lines. First, we filtered the variants using PLINK with the parameters '--indep-pairwise 50 10 0.2', leaving 282 381 unlinked variants. These variants were used for ADMIXTURE analysis, considering K values from $K = 1$ to $K = 10$. The optimal K value ($K = 4$) was calculated using the cross-validation error function included in ADMIXTURE (see code in Appendix S1).

After removing the outlier SNPs that were detected with pcadapt (see 'Genome scanning for selective sweep signals'), a PCA was performed and results were visualized with ggplot2 (Wickham, 2016) to explore patterns of genetic differentiation between samples. A relative dissimilarity distance matrix was calculated via the function bitwise.dist from poppr v2.9.2 (Kamvar et al., 2014) and used to construct an NJ dendrogram using ape v5.5 (Paradis & Schliep, 2019) and ggtree v3.0.2 (Yu et al., 2017).

For genetic structure analysis using SV data, genotypes were coded as haploid presence–absence data and inserted into R 4.1.0 using adegenet v2.1.3 (Jombart, 2008). PCAs based on Euclidean distances were conducted separately for each type of SV using the cmdscale function of the stats package (R Core Team, 2021). After confirming the emergence of a common pattern from the individual PCA plots, a new plot sourcing the entirety of the dataset was produced. A relative dissimilarity distance matrix was calculated via the function diss.dist from poppr and used to construct an NJ dendrogram as described for the SNP dataset. The code for the above analyses is provided in Appendix S1.

Nucleotide variability was estimated for the whole genome by sliding 20-kb windows along the chromosomes using the Watterson theta (Watterson, 1975) and Tajima's theta (Tajima, 1989). Tajima's D test of neutrality (Tajima, 1989) and F_{ST} (Hudson et al., 1992) were also calculated in the analyses using *mstatspop* (Guirao-Rico et al., 2018, available at https://github.com/cragenomica/mstatspop).

TreeMix (Pickrell & Pritchard, 2012) was used to build a maximum likelihood tree. The best value for migration events (m) was determined by executing 10 replicate runs per number of m and utilizing the 99.8% variance in relatedness between populations threshold as suggested by Pickrell and Pritchard (2012) via the OptM R package (Fitak, 2021).

Genome-wide association study

The list of phenotyped olive tree cultivars and their SNPs on chromosomes 1–23 were kept for the GWAS (Table **S2**). The different number of cultivars included in the three association panels was due to missing phenotypes for the respective cultivars. The SNP variants were filtered for MAF (≥0.1) (Atwell et al., 2010), minor allele number (≥5), and missing rate (≤0.5) using PLINK v1.9 (Chang et al., 2015; Purcell et al., 2007). Biallelic SNPs were quality checked using SNPrelate (v1.22.0) (Zheng et al., 2012). The EMMAX pipeline was employed for association tests (Kang et al., 2010). To correct for population structure, the kinship matrices were generated using the EMMAX-BN (Balding-Nichols) method and the Manhattan and QQ plots were visualized using R (R Core Team, 2022; Song et al., 2018). To correct for multiple testing, Bonferroni correction was used. The Bonferroni method was used to control the family-wise error rate.

ACKNOWLEDGMENTS

This research was financed by Greek Public Investments Program (PIP) of General Secretariat for Research & Technology (GSRT), under the Emblematic Action 'The Olive Road' (project code: 2018ΣΕ01300000). Sebastián Ramos-Onsins is supported by the grant PID2020-119255GB-I00 (MICINN, Spain) and the CERCA Programme/Generalitat de Catalunya and acknowledges financial support from the Spanish Ministry of Economy and Competitiveness, through the Severo Ochoa Programme for Centres of Excellence in R&D 2016–2019 and 2020–2023 (SEV-2015-0533, CEX2019-000917) and the European Regional Development Fund (ERDF). The publication of the article in OA mode was financially supported by HEAL-Link.

CONFLICT OF INTEREST

The authors have not declared a conflict of interest.

DATA AVAILABILITY STATEMENT

The raw sequencing data, in FASTQ format, have been deposited in the National Center for Biotechnology

Information (NCBI) Short Read Archive (SRA) database under accession number PRJNA782823.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Projection of principal components 3 and 4 for the 89 *O. europaea* genotypes based on SNPs.

Figure S2. Manhattan and quantile-quantile (QQ) plots of genome-wide association scan for (a, b) apex position, (c, d) base position, (e, f) endocarp position, (g, h) endocarp shape, (i, j) endocarp apex shape, (k, l) inflorescence length, (m, n) leaf width, (o, p) leaf length, (q, r) leaf shape, and (s, t) leaf curvature. Log transformed *P*-values are plotted against the physical position on the respective chromosomes for 428 320 SNPs with minor allele frequency > 0.1. In each plot genomic coordinates are displayed on the *x*-axis. The negative logarithm of the association *P*-value for every SNP is plotted on the *y*-axis, meaning that every dot on the Manhattan plot signifies a SNP. The horizontal red line indicates the Bonferroni threshold.

Figure S3. Manhattan and quantile-quantile (QQ) plots of genome-wide association scan for (a, b) endocarp symmetry, (c, d) fruit lenticels, (e, f) size of lenticels, (g, h) fruit symmetry, (i, j) fruit shape, (k, l) olive ripening time, (m, n) fruit nipple, (o, p) endocarp terminal apex, (q, r) endocarp base, (s, t) endocarp number of grooves, (u, v) endocarp weight, and (w, x) olive oil content. Log transformed *P*-values are plotted against the physical position on the respective chromosomes for 428 320 SNPs with minor allele frequency > 0.1. In each plot genomic coordinates are displayed on the *x*-axis. The negative logarithm of the association *P*-value for every SNP is plotted on the *y*-axis, meaning that every dot on the Manhattan plot signifies a SNP. The horizontal red line indicates the Bonferroni threshold.

Table S1. The origin of genotypes and data used in the present study.

Table S2. Phenotyping data of morphological and agronomic traits.

Table S3. Sequencing data generated for the 44 newly resequenced olive genotypes.

Table S4.1. List of CNVs in the 89 samples, grouped based on the CNV type and represented in a binary format (0: absence; 1: presence).

Table S4.2. List of CNVs in genic regions, grouped based on CNV type.

Table S4.3. Functional annotation of genes within CNVs.

Table S4.4. List of deletions and duplications in the 89 samples, represented in a binary format (0: absence; 1: presence).

Table S4.5. Functional annotation of genes containing deletions or duplications.

Table S5.1. Annotation of SNPs of the 89 genotypes.

Table S5.2. Annotation of SNPs of the NORTH-A MED population.

Table S5.3. Annotation of SNPs of the NORTH-B MED population.

Table S5.4. Annotation of SNPs of the NORTH-EAST MED population.

Table S5.5. Annotation of SNPs of the WILD population.

Table S5.6. Annotation of SNPs of the OUT population.

Table S6. Sanger sequencing results of 16 SNPs in 16 olive genotypes. The 0 allele corresponds to the reference SNP and the 1 allele corresponds to the alternative SNP, so 0/0 means homozygous for the reference SNP, 0/1 means heterozygous, and 1/1 means homozygous for the alternative SNP.

Table S7.1. Annotation of InDels of the 89 genotypes.

Table S7.2. Annotation of InDels of the NORTH-A MED population.

Table S7.3. Annotation of InDels of the NORTH-B MED population.

Table S7.4. Annotation of InDels of the NORTH-EAST MED population.

Table S7.5. Annotation of InDels of the WILD population.

Table S8. Position and functional annotations of selection sweep gene candidates.

Table S9. SNP variants in genes involved in flowering time, self-compatibility, and fruit weight.

Table S10. Annotated candidate genes within a 100-kb interval from the significant SNPs for every trait in GWAS.

Appendix S1. R and shell scripts used in the present study.

REFERENCES

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L. *et al.* (2020) Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell*, **182**, 145–161.e23. Available from: https://doi.org/10.1016/j.cell.2020.05.021

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., Willems, G., Horton, M., Li, Y. *et al.* (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*, **465**, 627–631.

Aukerman, M.J. & Sakai, H. (2003) Regulation of flowering time and floral organ identity by a microRNA and Its APETALA2-like target genes. *Plant Cell*, **15**, 2730–2741.

Aydin, M., Tombuloglu, H., Hernandez, P., Dorado, G. & Unver, T. (2021) Olive-tree genome sequencing: towards a better understanding of oil biosynthesis. In: Tombuloglu, H., Unver, T., Tombuloglu, G. & Hakeem, K.R. (Eds.) *Oil crop genomics. Cham: Springer Nature*, pp. 75–87. Available from: https://doi.org/10.1007/978-3-030-70420-9_4

Baali-Cherif, D. & 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**, 823–830.

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**, 776–791.

Barranco, D., Cimato, A., Fiorino, P., Rallo, L., Touzani, A., Castañeda, C. *et al.* (2000) *World catalogue of olive varieties*. Madrid, Spain: International Olive Oil Council.

Bartolini, G. & Petruccelli, R. (2002) *Classification, origin, diffusion and history of the olive*. Rome: Food & Agriculture Organization of the United Nations.

Bazakos, C., Hanemian, M., Trontin, C., Jimenez-Gomez, J.M. & Loudet, O. (2017) New strategies and tools in quantitative genetics: how to go from the phenotype to the genotype. *Annual Review of Plant Biology*, **68**, 435–455.

Belaj, A., Dominguez-García, M.D.C., Atienza, S.G., Urdíroz, N.M., De la Rosa, R., Satovic, Z. *et al.* (2012) Developing a core collection of olive (*Olea europaea* L.) based on molecular markers (DArTs, SSRs, SNPs) and agronomic traits. *Tree Genetics and Genomes*, **8**, 365–378. Available from: https://doi.org/10.1007/s11295-011-0447-6

Besnard, G., Hernández, P., Khadari, B., Dorado, G. & Savolainen, V. (2011) Genomic profiling of plastid DNA variation in the Mediterranean olive tree. *BMC Plant Biology*, **11**, 1–12. Available from: https://doi.org/10.1186/1471-2229-11-80

Besnard, G., Khadari, B., Navascués, M., Fernández-Mazuecos, M., el Bakkali, A., Arrigo, N. *et al.* (2013) The complex history of the olive tree: from late quaternary diversification of mediterranean lineages to primary domestication in the northern Levant. *Proceedings of the Royal Society B: Biological Sciences*, **280**, 20122833. Available from: https://doi.org/10.1098/rspb.2012.2833

Besnard, G., Terral, J.F. & Cornille, A. (2018) On the origins and domestication of the olive: a review and perspectives. *Annals of Botany*, **121**, 385–403.

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*, **28**, 423–425.

Byzova, M.V., **Franken, J.**, **Aarts, M.G.M.**, **De Almeida-Engler, J.**, **Engler, G.**, **Mariani, C.** ***et al.*** (1999) Arabidopsis STERILE APETALA, a multifunctional gene regulating inflorescence, flower, and ovule development. *Genes & Development*, **13**, 1002–1014.

Chang, C.C., **Chow, C.C.**, **Tellier, L.C.A.M.**, **Vattikuti, S.**, **Purcell, S.M.** & **Lee, J.J.** (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, **4**, 7.

Chen, X., **Schulz-Trieglaff, O.**, **Shaw, R.**, **Barnes, B.**, **Schlesinger, F.**, **Källberg, M.** ***et al.*** (2016) Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, **32**(8), 1220–1222. Available from: https://doi.org/10.1093/bioinformatics/btv710

Cingolani, P., **Platts, A.**, **Wang, L.L.**, **Coon, M.**, **Nguyen, T.**, **Wang, L.** ***et al.*** (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin).*, **6**, 80–92. Available from: https://doi.org/10.4161/fly.19695

Corominas-Faja, B., **Cuyàs, E.**, **Lozano-Sánchez, J.**, **Cufí, S.**, **Verdura, S.**, **Fernández-Arroyo, S.** ***et al.*** (2018) Extra-virgin olive oil contains a metabolo-epigenetic inhibitor of cancer stem cells. *Carcinogenesis*, **39**, 601–613.

Cruz, F., **Julca, I.**, **Gómez-Garrido, J.**, **Loska, D.**, **Marcet-Houben, M.**, **Cano, E.** ***et al.*** (2016) Genome sequence of the olive tree, *Olea europaea*. *Gigascience*, **5**, 29.

DiMario, R.J., **Machingura, M.C.**, **Waldrop, G.L.** & **Moroney, J.V.** (2018) The many types of carbonic anhydrases in photosynthetic organisms. *Plant Science*, **268**, 11–17.

Faize, M., **Fumanal, B.**, **Luque, F.**, **Ramírez-Tejero, J.A.**, **Zou, Z.**, **Qiao, X.** ***et al.*** (2020) Genome wild analysis and molecular understanding of the aquaporin diversity in olive trees (*Olea europaea* l.). *International Journal of Molecular Sciences*, **21**, 1–31.

Fitak, R.R. (2021) OptM: estimating the optimal number of migration edges on population trees using Treemix. *Biology Methods and Protocols*, **6**, bpab017.

Gros-Balthazard, M., **Besnard, G.**, **Sarah, G.**, **Holtz, Y.**, **Leclercq, J.**, **Santoni, S.** ***et al.*** (2019) Evolutionary transcriptomics reveals the origins of olives and the genomic changes associated with their domestication. *The Plant Journal*, **100**, 143–157. Available from: https://doi.org/10.1111/tpj.14435

Guirao-Rico, S., **Ramirez, O.**, **Ojeda, A.**, **Amills, M.** & **Ramos-Onsins, S.E.** (2018) Porcine Y-chromosome variation is consistent with the occurrence of paternal gene flow from non-Asian to Asian populations. *Heredity*, **120**(1), 63–76. Available from: https://doi.org/10.1038/s41437-017-0002-9

Haberman, A., **Bakhshian, O.**, **Cerezo-Medina, S.**, **Paltiel, J.**, **Adler, C.**, **Ben-Ari, G.** ***et al.*** (2017) A possible role for flowering locus T-encoding genes in interpreting environmental and internal cues affecting olive (*Olea europaea* L.) flower induction. *Plant, Cell & Environment*, **40**, 1263–1280. Available from: https://doi.org/10.1111/pce.12922

Harwood, L.M. & **Moody, C.J.** (1989) *Experimental organic chemistry: principles and Practice*. Hoboken,NJ: Blackwell Scientific.

Hudson, R.R., **Slatkin, M.** & **Maddison, W.P.** (1992) Estimation of levels of gene flow from DNA sequence data. *Genetics*, **132**(2), 583–589. Available from: https://doi.org/10.1093/genetics/132.2.583

Jeffares, D.C., **Jolly, C.**, **Hoti, M.**, **Speed, D.**, **Shaw, L.**, **Rallis, C.** ***et al.*** (2017) Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature Communications*, **8**, 1–11.

Jiménez-Ruiz, J., **Ramírez-Tejero, J.A.**, **Fernández-Pozo, N.**, **de la O Leyva-Pérez, M.**, **Yan, H.**, **de la Rosa, R.** ***et al.*** (2020) Transposon activation is a major driver in the genome evolution of cultivated olive trees (*Olea europaea* L.). *Plant Genome*, **13**, e20010. Available from: https://doi.org/10.1002/tpg2.20010

Jombart, T. (2008) Adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405.

Julca, I., **Marcet-Houben, M.**, **Cruz, F.**, **Gómez-Garrido, J.**, **Gaut, B.S.**, **Díez, C.M.** ***et al.*** (2020) Genomic evidence for recurrent genetic admixture during the domestication of Mediterranean olive trees (*Olea europaea* L.). *BMC Biology*, **18**, 1–25. Available from: https://doi.org/10.1186/s12915-020-00881-6

Kamvar, Z.N., **Tabima, J.F.** & **Grunwald, N.J.** (2014) Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, **2014**, 1–14.

Kang, H., **Zhang, T.T.**, **Li, Y.Y.**, **Lin-Wang, K.**, **Espley, R.V.**, **du, Y.P.** ***et al.*** (2022) The apple BTB protein MdBT2 positively regulates MdCOP1 abundance to repress anthocyanin biosynthesis. *Plant Physiology*, **190**, 305–318.

Kang, H.M., **Sul, J.H.**, **Service, S.K.**, **Zaitlen, N.A.**, **Kong, S.Y.**, **Freimer, N.B.** ***et al.*** (2010) Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, **42**, 348–354.

Kaya, H.B., **Akdemir, D.**, **Lozano, R.**, **Cetin, O.**, **Sozer Kaya, H.**, **Sahin, M.** ***et al.*** (2019) Genome wide association study of 5 agronomic traits in olive (*Olea europaea* L.). *Scientific Reports*, **9**, 1–14.

Kersey, P.J. (2019) Plant genome sequences: past, present, future. *Current Opinion in Plant Biology*, **48**, 1–8.

Koniditsiotis, S. (2020) Registration and promotion of monumental olive trees in Greece. *Advances in Social Sciences Research Journal*, **7**, 107–121.

Kostelenos, G. & **Kiritsakis, A.** (2017) Olive tree history and evolution. In: Kiritsakis, A. & Shahidi, F. (Eds.) *Olives olive oil as functional foods*. Hoboken, NJ: John Wiley & Sons, Inc, pp. 1–12. Available from: https://doi.org/10.1002/9781119135340.ch1

Koubouris, G.C., **Avramidou, E.V.**, **Metzidakis, I.T.**, **Petrakis, P.V.**, **Sergentani, C.K.** & **Doulis, A.G.** (2019) Phylogenetic and evolutionary applications of analyzing endocarp morphological characters by classification binary tree and leaves by SSR markers for the characterization of olive germplasm. *Tree Genetics & Genomes*, **15**, 1–12. Available from: https://doi.org/10.1007/s11295-019-1322-0

Layer, R.M., **Chiang, C.**, **Quinlan, A.R.** & **Hall, I.M.** (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology*, **15**, 1–19.

Li, Y., **Cao, K.**, **Zhu, G.**, **Fang, W.**, **Chen, C.**, **Wang, X.** ***et al.*** (2019) Genomic analyses of an extensive collection of wild and cultivated accessions provide new insights into peach breeding history. *Genome Biology*, **20**, 1–18. Available from: https://doi.org/10.1186/s13059-019-1648-9

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. Available from: https://doi.org/10.1093/bioinformatics/btp352

Loumou, A. & **Giourga, C.** (2003) Olive groves: "the life and identity of the Mediterranean". *Agriculture and Human Values*, **20**, 87–95. Available from: https://doi.org/10.1023/A:1022444005336

Lovisolo, C., **Secchi, F.**, **Nardini, A.**, **Salleo, S.**, **Buffa, R.** & **Schubert, A.** (2007) Expression of PIP1 and PIP2 aquaporins is enhanced in olive dwarf genotypes and is related to root and leaf hydraulic conductance. *Physiologia Plantarum*, **130**, 543–551. Available from: https://doi.org/10.1111/j.1399-3054.2007.00902.x

Manichaikul, A., **Mychaleckyj, J.C.**, **Rich, S.S.**, **Daly, K.**, **Sale, M.** & **Chen, W.M.** (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics*, **26**(22), 2867–2873.

Mariotti, R., **Fornasiero, A.**, **Mousavi, S.**, **Cultrera, N.G.M.**, **Brizioli, F.**, **Pandolfi, S.** ***et al.*** (2020) Genetic mapping of the incompatibility locus in olive and development of a linked sequence-tagged site marker. *Frontiers in Plant Science*, **10**, 1760.

Meyer, R.S. & **Purugganan, M.D.** (2013) Evolution of crop species: genetics of domestication and diversification. *Nature Reviews. Genetics*, **14**, 840–852.

Moret, M., **Ramírez-Tejero, J.A.**, **Serrano, A.**, **Ramírez-Yera, E.**, **Cueva-López, M.D.**, **Belaj, A.** ***et al.*** (2023) Identification of genetic markers and genes putatively involved in determining olive fruit weight. *Plants*, **12**, 155.

Pannelli, G., **Pandolfi, S.**, **Baldoni, L.** & **Bongi, G.** (2010) Selezione e valorizzazione di olivi antichi in Umbria. *Proceedings of the IV Convegno Nazionale Piante Mediterranee. Le potenzialità del territorio e dell'ambiente*, Marina di Nova Siri (MT), Italy, pp. 93–104.

Paradis, E. & **Schliep, K.** (2019) Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, **35**, 526–528.

Pickrell, J. & **Pritchard, J.** (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, **8**, e1002967.

Privé, F., **Luu, K.**, **Vilhjálmsson, B.J.**, **Blum, M.G.B.** & **Rosenberg, M.** (2020) Performing highly efficient genome scans for local adaptation with R package pcadapt version 4. *Molecular Biology and Evolution*, **37**, 2153–2154.

Purcell, S., **Neale, B.**, **Todd-Brown, K.**, **Thomas, L.**, **Ferreira, M.A.R.**, **Bender, D.** ***et al.*** (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, **81**, 559–575.

R Core Team. (2021) *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from: https://www.R-project.org/ [Accessed 17th May 2023].

R Core Team. (2022) *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from: https://www.R-project.org/ [Accessed 17th May 2023].

Rao, G., **Zhang, J.**, **Liu, X.**, **Lin, C.**, **Xin, H.**, **Xue, L.** ***et al.*** (2021) De novo assembly of a new *Olea europaea* genome accession using nanopore sequencing. *Horticulture Research*, **8**, 64. Available from: https://doi.org/10.1038/s41438-021-00498-y/6446608

Rausch, T., **Zichner, T.**, **Schlattl, A.**, **Stütz, A.M.**, **Benes, V.** & **Korbel, J.O.** (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**, 333–339.

Rimmer, A., **Phan, H.**, **Mathieson, I.**, **Iqbal, Z.**, **Twigg, S.R.F.**, **Wilkie, A.O.M.** ***et al.*** (2014) Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, **46**, 912–918.

Rong, S., **Wu, Z.**, **Cheng, Z.**, **Zhang, S.**, **Liu, H.** & **Huang, Q.** (2020) Genome-wide identification, evolutionary patterns, and expression analysis of bzip gene family in olive (*Olea europaea* L.). *Genes (Basel)*, **11**, 510.

Rosati, A., **Caporali, S.** & **Paoletti, A.** (2012) Floral biology: implications for fruit characteristics and yield. In: Muzzalupo, I. (Ed.) *Olive germplasm—the olive cultivation, table olive and olive oil industry in Italy*. London, UK: InTech, pp. 71–80.

Rubio De Casas, R., **Besnard, G.**, **Schönswetter, P.**, **Balaguer, L.** & **Vargas, P.** (2006) Extensive gene flow blurs phylogeographic but not phylogenetic signal in *Olea europaea* L. *Theoretical and Applied Genetics*, **113**, 575–583.

Rudenko, N.N., **Ignatova, L.K.**, **Nadeeva-Zhurikova, E.M.**, **Fedorchuk, T.P.**, **Ivanov, B.N.** & **Borisova-Mubarakshina, M.M.** (2021) Advances in understanding the physiological role and locations of carbonic anhydrases in C3 plant cells. *Protoplasma*, **258**, 249–262. Available from: https://doi.org/10.1007/s00709-020-01566-1

Rugini, E., **Baldoni, L.**, **Muleo, R.** & **Sebastiani, L.** (2016) *The olive tree genome*. Berlin, Germany: Springer International Publishing.

Sanz-Cortés, F., **Martinez-Calvo, J.**, **Badenes, M.L.**, **Bleiholder, H.**, **Hack, H.**, **Llacer, G.** ***et al.*** (2002) Phenological growth stages of olive trees (*Olea europaea*). *The Annals of Applied Biology*, **140**, 151–157. Available from: https://doi.org/10.1111/j.1744-7348.2002.tb00167.x

Sato, F. & **Matsui, K.** (2011) Engineering the biosynthesis of low molecular weight metabolites for quality traits (essential nutrients, health-promoting phytochemicals, volatiles, and aroma compounds). *Plant Biotechnology and Agriculture*, **2012**, 443–461.

Skodra, C., **Titeli, V.S.**, **Michailidis, M.**, **Bazakos, C.**, **Ganopoulos, I.**, **Molassiotis, A.** ***et al.*** (2021) Olive fruit development and ripening: break on through to the "-omics" side. *International Journal of Molecular Sciences*, **22**, 5806.

Song, B., **Mott, R.** & **Gan, X.** (2018) Recovery of novel association loci in *Arabidopsis thaliana* and *Drosophila melanogaster* through leveraging INDELs association and integrated burden test. *PLoS Genetics*, **14**, e1007699. Available from: https://doi.org/10.1371/journal.pgen.1007699

Storey, J.D., **Bass, A.J.**, **Dabney, A.** & **Robinson, D.** (2020) *qvalue: Q-value estimation for false discovery rate control*. R package version 2.20.0.

Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**(3), 585–595. Available from: https://doi.org/10.1093/genetics/123.3.585

Unver, T., **Wu, Z.**, **Sterck, L.**, **Turktas, M.**, **Lohaus, R.**, **Li, Z.** ***et al.*** (2017) Genome of wild olive and the evolution of oil biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America*, **114**, E9413–E9422. Available from: https://doi.org/10.1073/pnas.1708621114

Varshney, R.K., **Saxena, R.K.**, **Upadhyaya, H.D.**, **Khan, A.W.**, **Yu, Y.**, **Kim, C.** ***et al.*** (2017) Whole-genome resequencing of 292 pigeonpea accessions identifies genomic regions associated with domestication and agronomic traits. *Nature Genetics*, **49**, 1082–1088.

Vatansever, R., **Hernandez, P.**, **Escalante, F.J.**, **Dorado, G.** & **Unver, T.** (2022) Genome-wide exploration of oil biosynthesis genes in cultivated olive tree varieties (*Olea europaea*): insights into regulation of oil biosynthesis. *Functional & Integrative Genomics*, **22**, 171–178. Available from: https://doi.org/10.1007/s10142-021-00824-6

Vossen, P. (2007) Olive oil: history, production, and characteristics of the world's classic oils. *HortScience*, **42**, 1093–1100.

Wang, P., **Luo, Y.**, **Huang, J.**, **Gao, S.**, **Zhu, G.**, **Dang, Z.** ***et al.*** (2020) The genome evolution and domestication of tropical fruit mango. *Genome Biology*, **21**, 1–17. Available from: https://doi.org/10.1186/s13059-020-01959-8

Watterson, G.A. (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**(2), 256–276. Available from: https://doi.org/10.1016/0040-5809(75)90020-9

Wickham, H. (2016) *ggplot2: elegant graphics for data analysis*. New York: Springer-Verlag.

Xanthopoulou, A., **Ganopoulos, I.**, **Koubouris, G.**, **Tsaftaris, A.**, **Sergendani, C.**, **Kalivas, A.** ***et al.*** (2014) Microsatellite high-resolution melting (SSR-HRM) analysis for genotyping and molecular characterization of an *Olea europaea* germplasm collection. *Plant Genetic Resources*, **12**, 273–277.

Xanthopoulou, A., **Manioudaki, M.**, **Bazakos, C.**, **Kissoudis, C.**, **Farsakoglou, A.M.**, **Karagiannis, E.** ***et al.*** (2020) Whole genome re-sequencing of sweet cherry (*Prunus avium* L.) yields insights into genomic diversity of a fruit species. *Horticulture Research*, **7**, 60.

Yu, G., **Smith, D.K.**, **Zhu, H.**, **Guan, Y.** & **Lam, T.T.Y.** (2017) Ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, **8**, 28–36. Available from: https://doi.org/10.1111/2041-210X.12628

Zhang, M.Y., **Xue, C.**, **Hu, H.**, **Li, J.**, **Xue, Y.**, **Wang, R.** ***et al.*** (2021) Genome-wide association studies provide insights into the genetic determination of fruit traits of pear. *Nature Communications*, **12**, 1–10. Available from: https://doi.org/10.1038/s41467-021-21378-y

Zheng, X., **Levine, D.**, **Shen, J.**, **Gogarten, S.M.**, **Laurie, C.** & **Weir, B.S.** (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, **28**(24), 3326–3328. Available from: https://doi.org/10.1093/bioinformatics/bts606

Zhu, S., **Niu, E.**, **Shi, A.** & **Mou, B.** (2019) Genetic diversity analysis of olive germplasm (*Olea europaea* L.) with genotyping-by-sequencing technology. *Frontiers in Genetics*, **10**, 755.