

## **The genome of wild olive and the evolution of oil biosynthesis**

The wild olive genome and oil biosynthesis

**Turgay Unver<sup>a,1,2,3,4</sup>, Zhangyan Wu<sup>b,2</sup>, Lieven Sterck<sup>c,d</sup>, Mine Turktas<sup>e</sup>, Rolf Lohaus<sup>c,d</sup>, Zhen Li<sup>c,d</sup>, Ming Yang<sup>b</sup>, Lijuan He<sup>b</sup>, Tianquan Deng<sup>b</sup>, Francisco Javier Escalante<sup>f</sup>, Carlos Llorens<sup>g</sup>, Francisco J. Roig<sup>g</sup>, Iskender Parmaksiz<sup>h</sup>, Ekrem Dundar<sup>i</sup>, Fuliang Xie<sup>j</sup>, Baohong Zhang<sup>j</sup>, Arif Ipek<sup>e</sup>, Serkan Uranbey<sup>k</sup>, Mustafa Erayman<sup>l</sup>, Emre Ilhan<sup>l</sup>, Oussama Badad<sup>m</sup>, Hassan Ghazal<sup>n</sup>, David A. Lightfoot<sup>o</sup>, Pavan Kasarla<sup>o</sup>, Vincent Colantonio<sup>o</sup>, Huseyin Tombuloglu<sup>p</sup>, Pilar Hernandez<sup>q</sup>, Nurengin Mete<sup>r</sup>, Oznur Cetin<sup>r</sup>, Marc Van Montagu<sup>c,d,4</sup>, Huanming Yang<sup>b</sup>, Qiang Gao<sup>b</sup>, Gabriel Dorado<sup>s,1</sup>, Yves Van de Peer<sup>c,d,t,1,4</sup>**

<sup>a</sup>İzmir International Biomedicine and Genome Institute (iBG-izmir), Dokuz Eylül University, Inciralti, 35340 İzmir, Turkey; <sup>b</sup>BGI Shenzhen, 518038 Shenzhen, China; <sup>c</sup>Department of Plant Biotechnology and Bioinformatics, Ghent University, 9052 Ghent, Belgium; <sup>d</sup>Center for Plant Systems Biology, VIB, 9052 Ghent, Belgium; <sup>e</sup>Department of Biology, Faculty of Science, Cankiri Karatekin University, 18100 Cankiri, Turkey; <sup>f</sup>Plataforma de Genómica y Bioinformática de Andalucía (GBPA), 41013 Sevilla, Spain; <sup>g</sup>Biotechvana, 46980 Paterna (Valencia), Spain; <sup>h</sup>Department of Molecular Biology and Genetics, Faculty of Science, Gaziosmanpasa University, Tokat, Turkey; <sup>i</sup>Department of Molecular Biology and Genetics, Faculty of Science, Balikesir University, 10145 Balikesir, Turkey; <sup>j</sup>Department of Biology, East Carolina University, Greenville, NC 27858, USA; <sup>k</sup>Department of Field Crops, Faculty of Agriculture, Ankara University, Ankara, Turkey; <sup>l</sup>Department of Biology, Faculty of Arts and Science, Mustafa Kemal University, Hatay, Turkey; <sup>m</sup>Laboratory of Plant Physiology, University Mohamed V in Rabat, Morocco; <sup>n</sup>Polydisciplinary Faculty of Nador, University

Mohamed Premier, Nador, Morocco; <sup>o</sup>Department of Plant, Soil and Agricultural Systems, Southern Illinois University, Carbondale, IL 62901, USA; <sup>p</sup>Institute for Research and Medical Consultation, University of Dammam, 34212, Dammam, Saudi Arabia; <sup>q</sup>Instituto de Agricultura Sostenible (IAS-CSIC), 14004 Cordoba; <sup>r</sup>Olive Research Institute of Bornova, 35100, Izmir Turkey; <sup>s</sup>Departamento Bioquímica y Biología Molecular, Campus de Excelencia Internacional Agroalimentario (ceiA3), Universidad de Córdoba, 14071 Córdoba, Spain; <sup>t</sup>Department of Genetics, Genomics Research Institute, University of Pretoria, Pretoria 0028, South Africa.

## Footnotes

Author contributions: T.U., M.V.M., G.D., and Y.V.d.P. designed research and coordinated the project; Z.W., M.T., M.Y., L.H., T.D., I.P., A.I., S.U., M.E., E.I., N.M., H.Y., and Q.G., contributed to data production; T.U., Z.W., L.S., M.T., R.L., Z.L., M.Y., F.J.E., C.L., F.J.R., E.D., F.X., B.Z., O.B., H.G., D.A.L., P.K., V.C., H.T., P.H., N.M., O.C., G.D., and Y.V.d.P. performed research and analyzed data; T.U., L.S., R.L., G.D., and Y.V.d.P. wrote the manuscript.

The authors declare no conflict of interest.

<sup>1</sup>These authors contributed to the project leadership.

<sup>2</sup>These authors contributed equally to this work.

<sup>3</sup>Current address: Egitim Mah. Ekrem Guer Sok. No:26/3, 35340 Balcova, Izmir, Turkey.

<sup>4</sup>To whom correspondence may be addressed. E-mail: turgayunver@icloud.com; marc.vanmontagu@ugent.be; yves.vandepeer@psb.vib-ugent.be

## **Significance**

We sequenced the genome and transcriptomes of the wild olive (oleaster). More than 50,000 genes were predicted, and evidence was found for two relatively recent whole-genome duplication events, dated at about 28 and 59 million years ago. Whole genome sequencing, as well as gene expression studies, provide further insights into the evolution of oil biosynthesis, and will aid future studies aimed at further increasing the production of olive oil, which is a key ingredient of the healthy Mediterranean diet and has been granted a qualified health claim by FDA.

## Abstract

Here, we present the genome sequence and annotation of the wild olive tree (*Olea europaea* var. *sylvestris*), called oleaster, which is considered an ancestor of cultivated olive trees. More than 50,000 protein-coding genes were predicted, a majority of which could be anchored to 23 pseudo-chromosomes obtained through a newly constructed genetic map. The oleaster genome contains signatures of two Oleaceae-lineage specific paleopolyploidy events, dated at approximately 28 and 59 million years ago. These events contributed to the expansion and neofunctionalization of genes and gene families that play important roles in oil biosynthesis. The functional divergence of oil biosynthesis pathway genes, such as *FAD2*, *SACPD*, *EAR* and *ACPT*, following duplication, has been responsible for the differential accumulation of oleic and linoleic acids produced in olive compared to sesame, a closely related oil crop. Duplicated oleaster *FAD2* genes are regulated by a short-interfering RNA (siRNA) derived from a transposable element-rich region, leading to suppressed levels of *FAD2* gene expression. Additionally, neofunctionalization of members of the *SACPD* gene family has led to increased expression of *SACPD2*, 3, 5 and 7, consequently resulting in an increased desaturation of steric acid. Taken together, decreased *FAD2* expression and increased *SACPD* expression likely explain the accumulation of exceptionally high levels of oleic acid in olive. The oleaster genome thus provides important insights into the evolution of oil biosynthesis and will be a valuable resource for oil crop genomics.

/body

As a symbol of peace, fertility, health and longevity, the olive tree (*Olea europaea* L.) is a socio-economically important oil crop that is widely grown in the Mediterranean Basin. Belonging to the Oleaceae family (order Lamiales), it can biosynthesize essential unsaturated fatty acids and other important secondary metabolites, such as vitamins and phenolic compounds (1). The olive tree is a diploid ( $2n = 46$ ) allogamous crop that can be vegetatively propagated and live for thousands of years (2). Paleobotanical evidence suggests that olive oil was already produced in the Bronze Age (3). It has been thought that cultivated varieties were derived from the wild olive tree, called oleaster (*O. europaea* var. *sylvestris*), in Asia Minor, which then spread to Greece (4). Nevertheless, the exact domestication history of the olive tree is unknown (5). Due to their longevity, oleaster trees might be even related to Neolithic olive tree ancestors (2). Although the natural long generation time of olive trees has traditionally hindered breeding in this species, there are a few breeding programs involving sexual crosses that have generated interesting varieties for novel uses, like “Chiquitita”, specifically selected for high density hedgerow orchards (6).

The olive is tightly associated with the Mediterranean cuisine. However, its consumption also spread to America (United States, Mexico, Brazil, Argentina and Peru), Asia (China and India) and Australia. This expansion was, besides cultural, mainly due to the recognition of the beneficial dietetic properties of olive oil as a source of healthy fatty acids and micronutrients (antioxidants like phenolic compounds, including vitamin E, etc.). In fact, olive oil has been granted a qualified health claim, reducing cardiovascular disease incidence (coronary heart disease) (7), by the Food and Drug Administration (FDA) of the United States of America (USA) (<http://www.fda.gov>; Docket No. 2003Q-0559). As such, it represents the third FDA-

approved claim for conventional foods, after nuts and omega-3 fatty acids. Moreover, olive tree products and by-products are also being utilized for pharmaceutical and cosmetic purposes.

Traditionally, olive oil is obtained by pressing olive fruits. Olive fruits consist of 20 to 30% (w/w) oil, 17% cellulose, 4% carbohydrates, 2% protein and 0.1% micronutrients (1), with the rest (46.9 to 56.9%) being water. Both polyols (mannitol) and oligosaccharides (raffinose and stachyose) are synthesized in olive tree leaves, being further exported with sucrose into the fruits, for both general metabolism and as precursors of olive oil biosynthesis (8). Starting from a carbon source such as sucrose, long-chain fatty acids are synthesized, modified and degraded by the activity of enzymes, including fatty-acid synthases, elongases, desaturases and carboxylases (9). Fatty acids are the major constituent of triacylglycerols (TAG). In olive oil, TAG are mostly composed of monounsaturated oleic acid (C18:1) (~75% of all TAG), followed by saturated palmitic acid (C16) (~13.5%), polyunsaturated linoleic acid (c18:2  $\omega$ -6) (~5.5%) and  $\alpha$ -linolenic acid (c18:3  $\omega$ -3) (~0.75%) (10).

## Results

**Assembly of the oleaster genome.** The wild olive tree genome was shotgun sequenced (220x coverage), generating 515.7 Gbp of data (*SI Appendix*, Table S1). SOAPdenovo (11) was used to assemble the sequence reads, which resulted in a draft genome assembly of 1.48 Gbp, with scaffold N50 of 228 kbp (*SI Appendix*, Table S3, which is in agreement with genome size estimations from flow cytometry (*SI Appendix*, Fig. S1) and *k*-mer analysis (*SI Appendix*, Fig. S2a and *SI Appendix*, Table S2) (~1.46 Gbp). Using a newly constructed genetic map, 50% of sequences longer than 1 kbp (~572 Mbp) could be anchored into 23 linkage groups (Fig. 1 and Table 1).

**Genome annotation.** The annotation of the oleaster genome was carried out by combining three different approaches; namely, *ab initio* prediction, homology-based prediction and transcriptome mapping (Fig. 1; Table 1). About 51% of the genome assembly was found to be composed of repetitive DNA (Fig. 1), which is less than what was found for the draft genome of a recently published cultivated olive tree (63%) (12). Genome comparisons between oleaster and nine other plant species showed differences in gene numbers, transcript lengths and proportions of transposable elements (TEs; *SI Appendix*, Table S5b). TEs and interspersed repeats occupied ~43% of the genome (Table 1 and *SI Appendix*, Table S7). Long terminal-repeats (LTR) were the most abundant type of TE (40.3% of genome), which is in agreement with a previous analysis on a cultivated olive tree (38.8% of genome) (13), followed by DNA-type TE (4.6%; *SI Appendix*, Table S7). A total of 50,684 protein-coding genes were predicted on the current assembly, of which 47,124 genes (93%) were confirmed by RNA-seq data. Further, 31,245 genes were located on the anchored pseudo-chromosomes (Fig. 1, *SI Appendix*, Fig. S6 and *SI Appendix*, Tables S8–9).

About 90 million small RNA (sRNA) reads from six different tissues were used for non-coding RNA (ncRNA) annotation (*SI Appendix*, Tables S10–11 and *SI Appendix*, Figs. S8–9). A total of 498 conserved microRNA (miRNA) families and 125 novel miRNAs were identified. Considering highly conserved miRNAs and their function, 29,842 miRNA-target pairs, including 7,849 unique target genes, were predicted. A total of 4,606, 1,937 and 630 miRNA targets were associated with transcription factors, stress response genes and metabolism genes, respectively (*SI Appendix*, Table S12).

Oleaster protein-coding genes were functionally characterized through Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG), which allowed annotating 72.42% and 50.14% of all genes, respectively (*SI Appendix*, Table S13). KEGG metabolic pathway annotations of oleaster and eleven other plant species including other oil crops such as



*Sesamum indicum* (sesame) and *Glycine max* (soybean), as well as *Populus trichocarpa* (poplar) as a reference tree genome, *Utricularia gibba* (bladderwort) and *Mimulus guttatus* (monkey flower) as close relatives within the Lamiales and *Fraxinus excelsior* (European ash tree) as a member of the Oleaceae family, showed a majority of oleaster genes to be involved in folding, sorting, degradation (4,263), biosynthesis of secondary metabolites (2,236), carbohydrate metabolism (1,905), and lipid metabolism (811). Protein clustering of predicted oleaster genes with genes of other sequenced plant species resulted in 17,208 gene families, 1,070 of which were oleaster-specific and 7,522 were shared with the Lamiales *F. excelsior*, *S. indicum*, *M. guttatus* and *U. gibba*. Although the number of gene families is largely consistent across the different species, the oleaster genome contains a large number (8,986) of unique genes (*SI Appendix*, Fig. S11 and *SI Appendix*, Table S14).

**Genome evolution.** The oleaster genome contains multiple signatures of paleopolyploidy events. Distributions of synonymous substitutions per synonymous site ( $K_S$ ) for both the whole paranome (the set of all duplicated genes in the genome, *SI Appendix*, Fig. S12a) and duplicates retained in collinear regions only (i.e., excluding duplicates from small-scale duplications, *SI Appendix*, Fig. S12b) consistently showed two clear peaks of duplicates at  $K_S$  values around 0.25 and 0.75, respectively. Peaks at similar  $K_S$  values have been reported for duplicated genes in the genome of European ash (14) (*F. excelsior*, a sister to oleaster in Oleaceae). Most likely, these peaks indicate two rounds of ancient whole-genome duplication (WGD) in the oleaster lineage (15), shared by olive and ash (14). To establish the age of these two WGD, absolute phylogenomic dating (16) was carried out. Absolute dating suggests that the most recent WGD had occurred around 26 to 30 million years ago (Mya) (Fig. 2A) and the older one around 57 to 63 Mya (Fig. 2B). As with many other WGDs in different plant lineages, the latter event seems

to have occurred close to the K/Pg extinction event providing additional evidence that WGDs – at least in plants – might be linked with periods of environmental change or upheaval (17).

Paleopolyploidy events of similar age have been reported for other asterids in this period. Within the Solanales, a shared whole-genome triplication has been found in the lineage leading to *Solanum tuberosum* (potato) and *S. lycopersicum* (tomato), with an estimated age at around 57 to 65 Mya, using methods similar to the ones used here (16). Within the Lamiales, multiple WGD independent from the paleopolyploidy in the Solanales have been described: two or three in the lineage leading to *U. gibba* (one of which could be shared with *M. guttatus*) (18), and one in the lineage leading to *S. indicum* (estimated age similar to tomato) (19). This latter one and the oldest WGD in *U. gibba* could denote the same event, possibly even shared with the older WGD in the oleaster and ash lineage, or both could be independent ones, partly depending on their phylogenetic relationship (*SI Appendix*, section 3.2). Mean estimates for the divergence of oleaster from *S. indicum* are 69 to 74 Mya (20-22) or even older (23, 24) (Fig. 2C). Duplication and speciation events analyzed using fourfold synonymous third-codon transversion rates (4DTv) also showed that there were probably two WGDs in oleaster and one WGD in *S. indicum*, and that these likely occurred after their divergence (Fig. 2D). Thus, the above dates and 4DTv patterns suggest that both WGD events inferred from the oleaster genome (as well as from the ash one) are specific to Oleaceae and occurred independently of WGD in the lineage leading to *S. indicum*, *M. guttatus* and *U. gibba* (Fig. 2C) (see also (14)). This seems further supported by a phylogenomic analysis of duplicates from the older oleaster WGD, in which a majority of trees supported an Oleaceae-lineage-specific event (*SI Appendix*, section 3.4, *SI Appendix*, Table S15 and *SI Appendix*, Fig. S13). High colinearity among oleaster chromosomes forms additional evidence for WGDs. At least 78 duplicated homologous genomic segments, 12 of which being intrachromosomal, were identified in the oleaster

genome. Among them, chromosomes 1 and 12 (4,743 genes), 7 and 14 (2,300 genes) and 6 and 21 (1,361 genes) are remarkably collinear (*SI Appendix*, Table S16 and *SI Appendix*, Fig. S14).

**Evolutionary analysis of oil biosynthesis.** Olive oil is mainly composed of TAG formed by fatty acids (10). Here, genes involved in oil biosynthesis were annotated and grouped, according to their sequence identity, pathway and enzyme codes. KEGG pathway analysis of genes related to oil biosynthesis in oleaster and eleven other species showed that the oleaster genome has the highest fraction of pathways related to lipid metabolism and secondary metabolite biosynthesis. Out of 308 described pathway annotations, some of them, such as  $\text{Ca}^{2+}$ -transporting ATPase (K01537), acyl-CoA oxidase (K00232) and phosphatidylserine decarboxylase (K01613) are highly represented in the oleaster genome compared to others. To further compare the evolution of oil biosynthesis between oleaster and another major oil-bearing crop, oleaster and sesame genes were subjected to InParanoid ortholog analysis (25). Out of 2,327 oil biosynthesis genes in oleaster, 2,025 seem to have homologs in sesame. After excluding outparalogs, 911 groups of orthologs could be built, with 1,232 inparalogs for olive tree and 1,171 inparalogs from sesame. Interestingly, 563 oil biosynthesis genes showed a strict one-to-one orthology between oleaster and sesame (despite independent WGD in oleaster and sesame), while the rest of inparalogs (669 in oleaster and 608 in sesame) were the result of independent and lineage-specific duplication events (see also Fig. 2 C and D). Furthermore, 94 (35%) of 267 genes were found to be unique to oleaster, in comparison to sesame, in terms of oil biosynthesis metabolic pathway annotation. Comparing orthologous genes between oleaster and sesame showed that a large proportion of genes required for oil biosynthesis has been maintained as duplicated genes in the oleaster genome (1,962 genes in 221 families). In contrast, only a small number of gene families (54 genes in 27 families) showed contraction in oleaster.

Fatty-acid biosynthesis is one of the major steps of complex oil biosynthesis (26). It includes elongation, degradation and biosynthesis of unsaturated fatty acids and is carried out through the activity of a large number of genes encoding fatty acid synthases, elongases, desaturases and carboxylases. Although the poly-unsaturated fatty acid (PUFA) pathway is common in plants, and a considerable number of orthologous gene families (911, see above) are shared between oleaster and sesame, many important gene families involved in the oil biosynthesis pathway were found to be expanded in the oleaster genome, compared to sesame (Fig. 3; *SI Appendix*, Fig. S17). Besides the expansion of some oil biosynthesis gene families in the oleaster genome, the contraction of gene families encoding degrading/catabolic enzymes (such as dehydrogenases and hydrolases) may also be responsible for the differential fatty-acid accumulation in oleaster and sesame. For instance, the number of linoleic acid metabolism genes was found to be significantly smaller for oleaster (20) than for sesame (164).

To explore functional divergence following duplication, expression analyses were performed in different tissues, collected from ripe and unripe fruits. Interestingly, it was observed that the expression of duplicated oleaster fatty-acid desaturase (*FAD2*) genes (*FAD2-1*, *FAD2-2*, *FAD2-4* and *FAD2-5*) was downregulated in fruit tissues, especially during the lipid accumulation ripening stage. Suppression of the expression of these genes causes reduced desaturation of oleic acid into linoleic acid (Fig. 4 A–D). *FAD2* genes underwent at least two rounds of WGD events in oleaster, but only one duplication event in sesame (19) (Fig. 4 B–D). By mapping small RNA (sRNA) reads to 10 kbp regions encompassing the oleaster *FAD2* genes (*SI Appendix*, Fig. S26), we discovered that an siRNA, which originated from a transposable element-rich region (27), may bind specifically to the 5'-UTR region of duplicated copies of the *FAD2* gene transcripts, repressing expression in the fruit tissues. Due to the presence of an additional twelve nucleotides at the siRNA-binding site, the *FAD2-3* transcript, unlike the other *FAD2* transcripts, may not be regulated by the activity of the siRNA in ripe fruit (Fig. 5 and *SI*

Appendix, Fig. S27). The *FAD2-3* gene is actively expressed in fruits and is responsible for the conversion of only a relatively low amount of oleic acid into linoleic acid (Fig. 4B). Sesame seeds also showed a differential expression-pattern for *FAD2* genes (*FAD2-1* and *FAD2-2*), however with low diversity (*FAD2*,  $\pi = 0.0016$ ), as reported (19). Consequently, silencing effects caused by siRNA on *FAD2* olive gene-transcripts (*FAD2-1*, -2, -4 and -5) (Fig. 5), and the low diversity in *FAD2* genes of sesame (19), are likely responsible for the higher accumulation of oleic acid in oleaster, with respect to sesame.

Oleic acid as a major component of olive oil is formed by dehydrogenation from stearic acid by stearyl-ACP desaturase (*SACPD*), after which it is desaturated into linoleic acid by *FAD2* (7). Expression measurement of oleaster *SACPD* genes showed that *SACPD1* and *SACPD2* have upregulated expression in leaf tissues, while *SACPD7* is being highly expressed in fruit tissues. On the other hand, *SACPD5* was found to be overexpressed in stem and pedicel tissues. Additionally, expression patterns of *SACPD1*, 5 and 6 were found to be at relatively low levels in other tissues (Fig. 4B).

It appears that the oleaster key genes involved in the PUFA pathway such as enoyl ACP reductase (*EAR*),  $\beta$ -ketoacyl-ACP synthase II (*KASII*), beta-ketoacyl-ACP reductase (*FabG*), acyl carrier protein (ACP)-hydrolase/thioesterase (*ACPTE*), *SACPD* and *FAD2-1* have been expanded by WGD and/or segmental duplications (*SI Appendix*, Table S17 and *SI Appendix*, Figs. S28 and S29). Synteny analysis suggests that oleaster *FAD2-1/-2* and *SACPD6/7* paralogs have been duplicated through WGD (*SI Appendix*, Fig. S29A). Furthermore, *EAR* (52 genes), *ACPTE* (9 genes), *FabG* (34 genes) and *KASII* (7 genes) were shown to be expanded by WGD, tandem and segmental duplications (*SI Appendix*, Fig. S28 and *SI Appendix*, Fig. S29 B-E) and now have different expression patterns (Fig. 3 and 4).

## Discussion

To date, besides the wild olive tree, the sequencing and assembly of two cultivated olive tree genomes have been reported, namely *O. europaea* cv. Leccino (13) and *O. europaea* cv. Farga (12), at ~4x and ~150x coverage, respectively. The latter, with a size of 1.31 Gbp, was preliminary annotated solely by utilizing RNA sequencing (RNA-seq) data, which resulted in more than 56,000 protein-coding genes (12). Compared to the oleaster genome presented here, the cultivated olive tree has a smaller genome size, albeit with a higher number of genes. Unlike some previous reports on olive tree genome data, which lacked chromosome anchoring and genome-wide functional annotation (12, 13), our study includes a near-complete representation and localization of genes, repeat elements and sRNA, as well as functional and metabolic annotations and an evolutionary analysis of oil biosynthesis genes.

Absolute dating of the two identified WGD events in oleaster and 4DTv patterns suggest that both WGDs, which seem to be shared with the ash tree, are specific to Oleaceae and independent from WGDs reported in other non-Oleaceae Lamiales, including *S. indicum* (sesame) (Fig. 2C). This is also consistent with synteny results from the ash tree genome (14). The age of the older WGD is close to the Cretaceous–Paleogene (K/Pg) boundary. Additional Oleaceae genomes will be required to determine which of the other lineages within Oleaceae, apart from ash, share either of the two WGD, and whether one or both are related to patterns of diversification within the family (28).

Both the expansion of gene families, and the functional divergence of genes playing important roles in oil biosynthesis, may explain the higher accumulation of oleic acid (~75% of olive oil) rather than linoleic acid (~5.5% of olive oil) in oleaster (10). In sesame seed oil, both types of fatty acids are more evenly present (~40%) with lower variation ( $\pm 5\%$ ) (19, 29) (Fig. 3 and 4A). Due to gene expansion and loss events in oleaster with respect to the PUFA pathway

genes responsible for the accumulation of oleic and linoleic acids, the fatty-acid content of olive oil greatly differs from sesame seed oil (Fig. 4A) (10, 19).

Here, consistent with a previous report (27), we also describe an siRNA sequence that originated from a transposable-element-rich genomic region. To inhibit expression of duplicated copies of *FAD2* gene transcripts, this regulatory siRNA may specifically bind to the 5'-UTR region of the transcripts in fruit tissues during the oil production period. In a previous study (30), it was reported that mutations associated with a duplication of the *Oleate Desaturase* (*OD*) gene caused its silencing by binding of an siRNA, further promoting accumulation of high-levels of oleic acid in sunflower seeds. Similarly, suppression of *FAD2* gene expression as a result of gene expansion probably leads to the high oleic-acid content in oleaster.

Based on expression analysis, *SACPD6/7* may have evolved through subfunctionalization or neofunctionalization events, following their duplication (Fig. 4B). On the other hand, *FAD2-1/-2* have probably retained similar functions, as their expression patterns have not changed (Fig. 4B). Compared to sesame, expansion of *SACPD* genes (*SACPD1-7*) in oleaster has likely led to increased desaturation activity and increased expression, through neofunctionalization of *SACPD2*, 3, 5 and 7 in fruit and stem tissues (Fig. 4B). Thus, neofunctionalized *SACPD* gene copies in oleaster are likely also responsible for the differences in oleic- and linoleic-acid contents of olive and sesame (19, 30). Recently, it was observed that mutations in the soybean *SACPD-C* gene promote higher accumulation of leaf stearic-acid content, as well as changes in leaf structure and morphology (31). Therefore, *SACPD1* and 2, which are highly expressed in leaves, might be related with leaf morphology as well as oleic-acid accumulation in fruit with overexpressed levels of *SACPD7* (Fig. 4B).

## Methods

A full description of the methods can be found in the *SI Appendix*.

**Plant material.** A wild olive tree ( $2n = 46$ ) was selected for whole-genome shotgun and transcriptome sequencing. Genomic DNA was isolated from leaf tissue (32).

**Genome and transcriptome sequencing.** Sequencing libraries were prepared and sequenced on the Illumina HiSeq 2000 platform, followed by assembly with SOAPdenovo (11). Transcriptome libraries of four tissues including leaf, stem, pedicel and fruit (ripe and unripe), collected from two different seasons, were also sequenced.

**Genome assembly.** All sequence reads were assembled with the SOAPdenovo software (11, 33) producing a reference sequence of the oleaster genome. A total of 319.39 Gbp of clean data were assembled into contigs and scaffolds, using the de Bruijn graph-based assembler of SOAPdenovo with the following four steps: (i) building contigs and scaffolds; (ii) filling gaps, (iii) removing redundancy; and (iv) reconstructing scaffolds.

**Genetic map construction and chromosome anchoring.** DNA samples of each F1 individual and parents were digested with *Pst*I-*Mse*I restriction enzymes and then ligated with enzyme-compatible adapters. To increase the number of *Pst*I-*Mse*I fragments, PCR amplifications were performed as described (34). The DArTseq (35) genotyping-by-sequencing (GBS) approach was used to identify single-nucleotide polymorphisms (SNP). GBS data were analyzed using regression-mapping algorithm of JoinMap 4.0 software from Kyazma (Wageningen, Netherlands) to enable linkage-map construction. MapChart 2.0 (36) was used for the graphical presentation of linkage maps. Genetic linkage maps were constructed to develop the integrated genome map for anchoring the scaffolds, using 94 individuals from a cross-pollinated (CP)



population of a cross between cultivars Memecik and Uslu. For chromosome-scale pseudomolecule construction, two maps were established from two progenies: both F1 progenies of 92 individuals (Memecik  $\times$  Uslu). An integrated map including 1,307 markers was established (37), based on double heterozygous loci (38, 39). Genetic markers were mapped onto the scaffolds using the Burrows-Wheeler Aligner (BWA) software module for alignment (BWA aln) (40) with default parameters. Afterwards, anchoring of assembled scaffolds to genetic maps was achieved by applying the ALLMAPS software (41).

**Repeat element analyses.** Both homology-based and *de novo* approaches were used to find TEs in the oleaster genome. The homology-based approach involved applying commonly used databases of known repetitive-sequences, along with programs such as RepeatProteinMask and RepeatMasker (42). RepeatModeler <<http://www.repeatmasker.org/RepeatModeler.html>> was utilized with two *ab initio* repeat-prediction programs (RECON and RepeatScout) to identify repeat-element boundaries and family relationships among sequences. Tandem repeats were also searched for in the genome, using Tandem Repeats Finder (43).

**Gene prediction.** Homology-based and *de novo* methods, as well as RNA-seq data, were used to predict genes in the *O. europaea* var. *sylvestris* genome. GLEAN (44) was used to consolidate results. Protein sequences of *Arabidopsis thaliana*, *S. indicum*, *S. tuberosum* and *Vitis vinifera* were aligned with TBLASTN and genBLASTA (45) against the matching genomic sequence using GeneWise (46) for accurate spliced alignments. Next, the *de novo* gene-prediction methods GlimmerHMM (47) <<https://ccb.jhu.edu/software/glimmerhmm>> and Augustus (48) were used to predict protein-coding genes, with parameters trained for *O. europaea* var. *sylvestris*, *A. thaliana*, *S. indicum*, *S. tuberosum* and *V. vinifera*.

**Genome annotation.** Functional annotation was achieved by comparing predicted proteins against public databases, including UniProtKB/Swiss-Prot, KEGG and InterPro. Results are available online at the “Olive Genome Browser” <<http://olivegenome.org>> and ORCAE <<http://bioinformatics.psb.ugent.be/orcae>> Gene-family clustering was performed by OrthoMCL (49).

**Evolutionary analyses.** The GTR+gamma evolutionary model was applied to reconstruct a phylogenetic tree using 231 single-copy orthologous genes from 12 different plant genomes.  $K_S$ -based age distributions of oleaster were also constructed to unveil whole-genome duplication events in oleaster (15). Absolute dating of two identified WGD events in the oleaster genome was performed as previously described (*SI Appendix* section 3) (16). SyMAP (50) was used to identify synteny with other species (i.e., *S. indicum*, *V. vinifera*, *P. trichocarpa* and *S. tuberosum*). Circos (51) was applied to generate a circular visualization of the oleaster genome features. InParanoid was used to identify orthologs and paralogs with sesame, involved in the oil biosynthesis pathways. See *SI Appendix* section 3 for additional information.

**Availability of data.** The oleaster genome assembly has been deposited at National Center for Biotechnology Information (NCBI) GenBank <<https://www.ncbi.nlm.nih.gov/genbank>> (no: SUB2036311; BioProject ID: PRJNA350614). Transcriptome datasets were deposited at NCBI Sequence Read Archive (SRA) <<https://www.ncbi.nlm.nih.gov/sra>> (accessions SRR4473639, SRR4473641, SRR44742, SRR4473643, SRR4473644, SRR4473645, SRR4473646 and SRR4473647). The genome and annotation files were also uploaded into ORCAE <<http://bioinformatics.psb.ugent.be/orcae>>, Phytozome <<https://phytozome.jgi.doe.gov>> and the olive genome consortium website <<http://olivegenome.org>>. Correspondence and requests should be addressed to <[turgayunver@icloud.com](mailto:turgayunver@icloud.com)>.

**Conflict of interest.** We declare no competing financial interests.

**ACKNOWLEDGEMENTS.** We would like to thank Prof. Dr. Ali Ibrahim Savas (former rector of Cankiri Karatekin University) for his great support to sustain the genome-sequencing project. We are grateful to Prof. Dr. Mehmet Ozturk (Dokuz Eylul University) for critical reading of the manuscript. This project was funded by Cankiri Karatekin University, BAP (2012-10, FF12035L19), State Planning Organization of Turkey (DPT2010K120720), Ankara University, BAP (project no: 14B0447004); Mustafa Kemal University, BAP (project no: 12022); Gaziosman Pasa University, BAP (2013/27); Turkish Academy of Science (TUBA, GEBIP); Ministry of Food, Agriculture and Livestock of Turkey (TAGEM/BBAD/12/A08/P06/3); “Consejería de Agricultura y Pesca” (041/C/2007, 75/C/2009 and 56/C/2010), “Grupo PAI” (AGR-248) of “Junta de Andalucía” and “Universidad de Córdoba” (“Ayuda a Grupos”), Spain; the Multidisciplinary Research Partnership ‘Bioinformatics: from nucleotides to networks’ Project (no. 01MR0310W) of Ghent University and funding from the European Union Seventh Framework Program (FP7/2007-2013) under European Research Council Advanced Grant Agreement 322739 – DOUBLEUP.

## References

1. Tripoli E, et al. (2005) The phenolic compounds of olive oil: structure, biological activity and beneficial effects on human health. *Nutr Res Rev* 18(1):98-112.
2. Lumaret R, Ouazzani N (2001) Plant genetics. Ancient wild olives in Mediterranean forests. *Nature* 413(6857):700.
3. Riley FR (2002) Olive oil production on bronze age Crete: nutritional properties, processing methods and storage life of Minoan olive oil. *Oxford J Archaeol* 21(1):63-75.

4. de Candolle A (1883) *Origine des plantes cultivées* (Librairie Germer Baillière et C<sup>ie</sup>, Paris, France).
5. Diez CM, et al.. (2015) Olive domestication and diversification in the Mediterranean Basin. *New Phytol* 206(1):436-447.
6. Rallo L, Barranco D, de la Rosa R, León L (2008) ‘Chiquitita’olive. *HortScience* 43(2):529-531.
7. Estruch R, et al. (2013) Primary prevention of cardiovascular disease with a Mediterranean diet. *N Engl J Med* 368(14):1279-1290.
8. Conde C, Delrot S, Geros H (2008) Physiological, biochemical and molecular changes occurring during olive development and ripening. *J Plant Physiol* 165(15):1545-1562.
9. Bates PD, Stymne S, Ohlrogge J (2013) Biochemical pathways in seed oil synthesis. *Curr Opin Plant Biol* 16(3):358-364.
10. Rueda A, Seiquer I, Olalla M, Giménez R, Lara L, and Cabrera-Vique C (2014) Characterization of fatty acid profile of argan oil and other edible vegetable oils by gas chromatography and discriminant analysis. *J Chemistry* 2014:843908.
11. Li R, et al. (2010) The sequence and de novo assembly of the giant panda genome. *Nature* 463(7279):311-317.
12. Cruz F, et al. (2016) Genome sequence of the olive tree, *Olea europaea*. *Gigascience* 5:29.
13. Barghini E, et al. (2014) The peculiar landscape of repetitive sequences in the olive (*Olea europaea* L.) genome. *Genome Biol Evol* 6(4):776-791.
14. Sollars ES, et al. (2017) Genome sequence and genetic diversity of European ash trees. *Nature* 541(7636):212-216.
15. Vanneste K, Van de Peer Y, Maere S (2013) Inference of genome duplications from age distributions revisited. *Mol Biol Evol* 30(1):177-190.

16. Vanneste K, Baele G, Maere S, Van de Peer Y (2014) Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. *Genome Res* 24(8):1334-1347.
17. Van de Peer Y, Mizrachi E, Marchal K (2017) The evolutionary significance of polyploidy. *Nat Rev Genet*: in press (doi: doi:10.1038/nrg.2017.26).
18. Ibarra-Laclette E, et al. (2013) Architecture and evolution of a minute plant genome. *Nature* 498(7452):94-98.
19. Wang L, et al. (2014) Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis. *Genome Biol* 15(2):R39.
20. Bell CD, Soltis DE, Soltis PS (2010) The age and diversification of the angiosperms revisited. *Am J Bot* 97(8):1296-1303.
21. Magallón S, Gómez-Acevedo S, Sánchez-Reyes LL, Hernández-Hernández T (2015) A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytol* 207(2):437-453.
22. Yi D-K, Kim K-J (2012) Complete chloroplast genome sequences of important oilseed crop *Sesamum indicum* L. *PLoS ONE* 7(5):e35872.
23. Bremer K, Friis EM, Bremer B (2004) Molecular phylogenetic dating of asterid flowering plants shows early Cretaceous diversification. *Syst Biol* 53(3):496-505.
24. Wikström N, Kainulainen K, Razafimandimbison SG, Smedmark JE, Bremer B (2015) A revised time tree of the asterids: establishing a temporal framework for evolutionary studies of the coffee family (Rubiaceae). *PLoS ONE* 10(5):e0126690 (Erratum: *PLoS ONE* 11(6):e0157206).
25. Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* 314(5):1041-1052.

26. Harwood JL, Guschina IA (2013) Regulation of lipid synthesis in oil crops. *FEBS Lett* 587(13):2079-2081.
27. Kuang HH, et al. (2009) Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the Solanaceae: New functional implications for MITes. *Genome Res* 19(1):42-56.
28. Besnard G, Rubio de Casas R, Christin PA, Vargas P (2009) Phylogenetics of Olea (Oleaceae) based on plastid and nuclear ribosomal DNA sequences: tertiary climatic shifts and lineage differentiation times. *Ann Bot* 104(1):143-160.
29. Wei WL, et al. (2013) association analysis for quality traits in a diverse panel of Chinese sesame (*Sesamum indicum* L.) germplasm. *J Integr Plant Biol* 55(8):745-758.
30. Lacombe S, Souyris I, Berville AJ (2009) An insertion of oleate desaturase homologous sequence silences via siRNA the functional gene leading to high oleic acid content in sunflower seed oil. *Mol Genet Genomics* 281(1):43-54.
31. Lakhssassi N, et al. (2017) Stearoyl-acyl carrier protein desaturase mutations uncover an impact of stearic acid in leaf and nodule structure. *Plant Physiol* 174(3):1531-1543.
32. Sahu SK, Thangaraj M, Kathiresan K (2012) DNA Extraction protocol for plants with high levels of secondary metabolites and polysaccharides without using liquid nitrogen and phenol. *ISRN Mol Biol* 2012:205049.
33. Li R, et al. (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* 20(2):265-272.
34. Raman H, et al. (2014) Genome-wide delineation of natural variation for pod shatter resistance in *Brassica napus*. *PLoS One* 9(7):e101673.
35. Elshire RJ, et al. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6(5):e19379.

36. Voorrips RE (2002) MapChart: software for the graphical presentation of linkage maps and QTLs. *J Hered* 93(1):77-78.
37. Risterucci A, et al. (2000) A high-density linkage map of *Theobroma cacao* L. *Theor Appl Genet* 101(5-6):948-955.
38. Pugh T, et al. (2004) A new cacao linkage map based on codominant markers: development and integration of 201 new microsatellite markers. *Theor Appl Genet* 108(6):1151-1161.
39. Fouet O, et al. (2011) Structural characterization and mapping of functional EST-SSR markers in *Theobroma cacao*. *Tree Genet Genomes* 7(4):799-817.
40. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26(5):589-595.
41. Tang H, et al. (2015) ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol* 16:3.
42. Tarailo-Graovac M, Chen N (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* 25:4.10.1-4.10.14.
43. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27(2):573-580.
44. Elsik CG, et al. (2007) Creating a honey bee consensus gene set. *Genome Biol* 8(1):R13.
45. She R, Chu JS, Wang K, Pei J, Chen N (2009) GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res* 19(1):143-149.
46. Birney E, Clamp M, Durbin R (2004) GeneWise and genomewise. *Genome Res* 14(5):988-995.
47. Majoros WH, Pertea M, Salzberg SL (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20(16):2878-2879.

48. Stanke M, et al. (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* 34 (Suppl 2):W435-W439.
49. Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 13(9):2178-2189.
50. Soderlund C, Bomhoff M, Nelson WM (2011) SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res* 39(10):e68.
51. Krzywinski M, et al. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res* 19(9):1639-1645.



## Figure Legends

**Figure 1. The genomic landscape of oleaster.** The outer layer represents the karyotype ideogram (colored blocks), with minor and major ticks labeling each 5 Mbp and 25 Mbp, respectively. Genome features across the 23 chromosomes (distinct characters shown as different colors, as indicated in the legend). Gene density per Mbp. Gene expression patterns in average RPKM (range of RPKM values plotted from 0 to >1,000). Tandem duplication density per Mbp. (E) Percentage heatmap of repeat coverage per Mbp. Percentage of transposable elements per Mbp (ranges of values plotted from 0 to >50). Inner circular representation, showing interchromosomal synteny.

**Figure 2. Oleaster genome evolution.** (A) and (B) Phylogenomic dating of *O. europaea* var. *sylvestris* paralogues. Absolute age distribution for the most recent WGD event ( $K_S$  of about 0.25; *SI Appendix*, Fig. S12a), with a consensus WGD age estimate of 28 Mya, and 90% confidence interval of 26 to 30 Mya (A). Likewise, for the older WGD event ( $K_S$  of about 0.75; *SI Appendix*, Fig. S12b), with a consensus WGD age estimate of 59 Mya, and 90% confidence interval of 57 to 63 Mya (B). The solid black line represents the KDE of dated paralogues, and the vertical dashed black line corresponds to its peak, which was used as the consensus WGD age estimate. Grey lines represent density estimates from 2,500 bootstrap replicates, whereas vertical black dotted lines indicate the corresponding 90% confidence interval for the WGD age estimate. Blue histogram shows the raw distribution of dated paralogues. (C) Estimation of divergence time. Blue numbers on the nodes are divergence time to present (Mya). The two Oleaceae WGD are indicated on the tree (blue rectangles), as well as other known WGD described in the literature for the species shown (gray rectangles; faded out rectangles indicate that an absolute date has not been estimated). Note discussion of phylogenetic relationships in

*SI Appendix*, section 3.2. (D) Fourfold degenerate (4DTv) distributions for *S. indicum*, *V. vinifera* and *O. europaea* var. *sylvestris*. Abscissa and ordinate represent 4DTv distance (using the HKY model) and percentage of homologous gene pairs, respectively.

**Figure 3. Oleic-acid biosynthesis pathway in oleaster.** Genes involved in oleic-acid biosynthesis with their differential expression patterns in stem, leaf, pedicel and fruit tissues are shown. Heatmap data correspond to starting (July, J) and end (November, N) time points for olive oil biosynthesis. The first step of such biosynthesis is catalyzed by Acetyl-CoA carboxylase (ACC), carboxylating Acetyl-CoA to form malonyl-CoA, which is converted to malonyl-ACP by SMT. Malonyl-ACP first reacts with 3-keto acyl-ACP, which is elongated by six reaction cycles, where chain-extender units are added. Then, fatty-acid synthases (FAS) act on that substrate to produce saturated fatty-acid16 carbon palmitate, which will be desaturated to form unsaturated fatty acids, such as oleic acid in oleaster. ACPTE, acyl carrier protein (ACP)-hydrolase/thioesterase; BCCP, biotin carboxyl carrier-protein; EAR, enoyl-ACP reductase; Exp, expanded; F, fruit; FabG, beta-ketoacyl-ACP reductase; FabZ, beta-hydroxyacyl-ACP dehydrase; FAD, fatty-acid desaturase; KAS,  $\beta$ -ketoacyl-ACP synthase; L, leaf; P, pedicel; PUFA, polyunsaturated fatty-acid; S, stem; SACPD, stearoyl-ACP desaturase; and SMT, S-malonyltransferase. Sesame expression data retrieved from Sesame Functional Genomics Database (SesameFG) (<http://202.127.18.220/hg>).

**Figure 4. Oleic-acid biosynthesis pathway in oleaster.**

(A) Oil content of oleaster and sesame, with major genes involved in oil biosynthesis. (B) Heatmap analyses of oleaster and sesame *FAD2* and *SACPD* genes. Blue lines indicate paralogs, which share orthologs with sesame. The arrow represents upregulation of *FAD2-3* gene, compared to other paralogs, in July unripe and November ripe fruits. Genes with green

font color indicate unique genes in the wild olive tree, which have no orthologous counterpart in sesame, whereas red font color represents orthologous genes. Sesame genes were labeled with turquoise color. (C) and (D) Phylogenetic trees, showing the duplication history of sesame and oleaster *FAD2* and *SACPD* genes. Blue squares show duplicated genes after WGD and tandem duplications (see also *SI Appendix*, Fig. S28A and Table S31). DAP, days after pollination;

**Figure 5. Regulation of *FAD2* gene-expression by siRNA.** Possible siRNA-binding sites are marked on 5'-UTR regions. Interestingly, siRNA can bind to *FAD2-1*, *FAD2-2*, *FAD2-4* and *FAD2-5* transcripts but cannot bind to *FAD2-3* transcripts, due to the presence of 12 additional nucleotides in the binding site (see *SI Appendix*, Fig. S27). Red lines show siRNA molecules. CDS, coding sequence; UTR, untranslated region.

**Table 1. Statistics of the wild olive tree genome assembly and annotation**

<b>Genome</b>					
Size (n, Gbp)					1.48
Karyotype (chromosomes, 2n)					46=2n
GC content (percentage, with N/without N)					36.8/38.8
High-copy repeat no.	# LTR/Gypsy and Copia				1,182,454
	# LINE				43,834
	# DNA transposable element				219,901
	# Unknown				42,630
Gene					50,684
<b>Assembly</b>					
# Scaffold >100 bp/>1 kbp					2,356,597/42,843
N50>100 bp/>1 kbp					228.62/364.6
<b>Annotation</b>					
	Number	Total Size (Kbp)	Avg. size (bp)	Max. size (bp)	Min. size (bp)
mRNA	50,684	65,933.6	1300.9	48,863	99
CDS	50,684	52,756.9	1,040.9	16,602	99
Exon	235,149	65,933.6	223.4	7,913	1
Intron	184,465	87,396.5	473.8	42,191	10
miRNA	411	49.979	113.33	24	21
tRNA	798	59.716	74.83	95	63
rRNA	773	121.906	121	1,804	29
snRNA	422	47.737	113	217	62
Tandem repeat	454,960	372,874.8	819.57	500,000	25
TE protein	428,172	23,958.1	559.54	5,505	24
Transposon	320,201	150,867.9	471.16	5,928	11
5'-UTR	15,172	8,002.1	527.42	38,088	5
3'-UTR	15,075	7,337	486.7	47,263	5

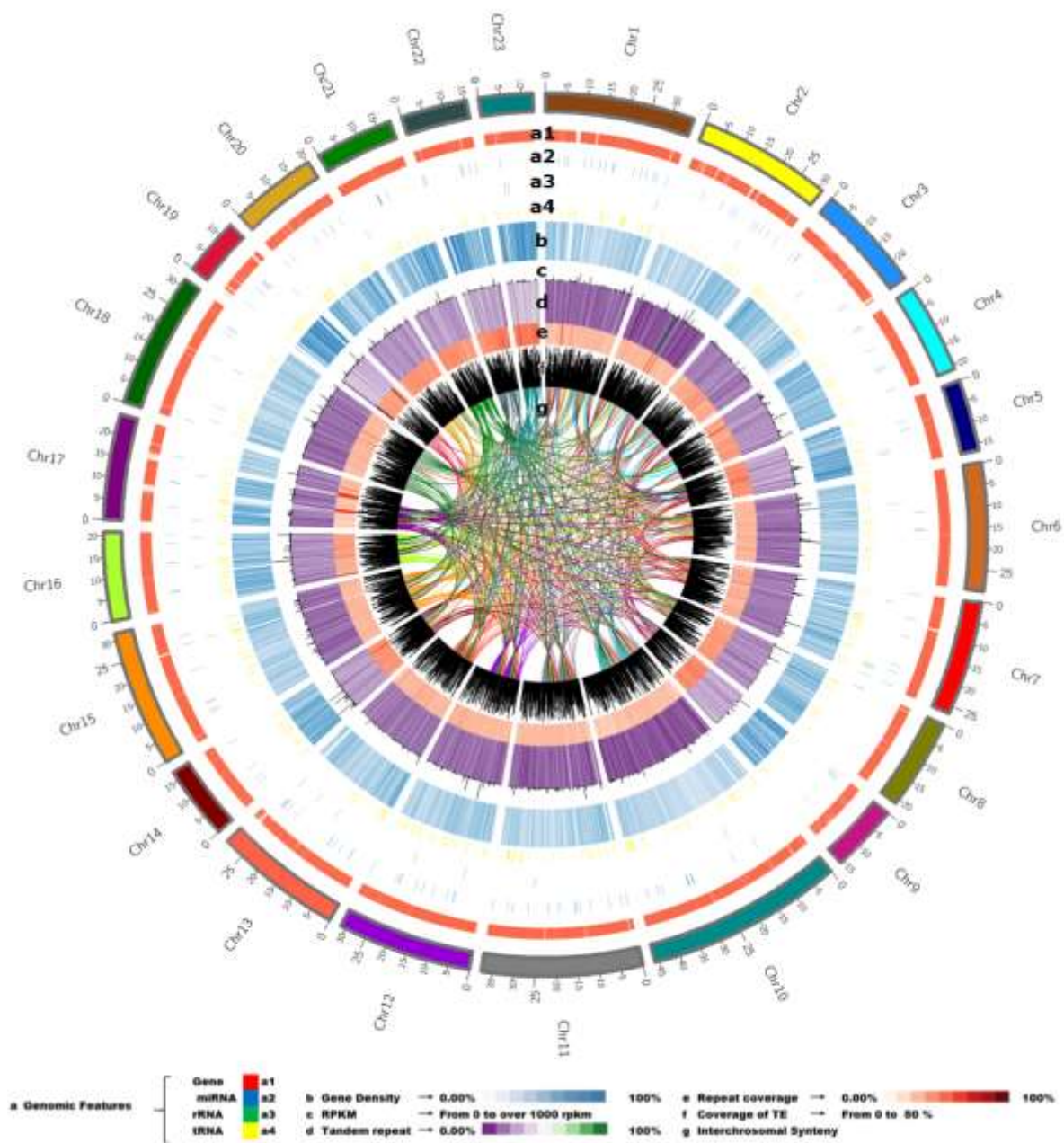


Fig. 1.

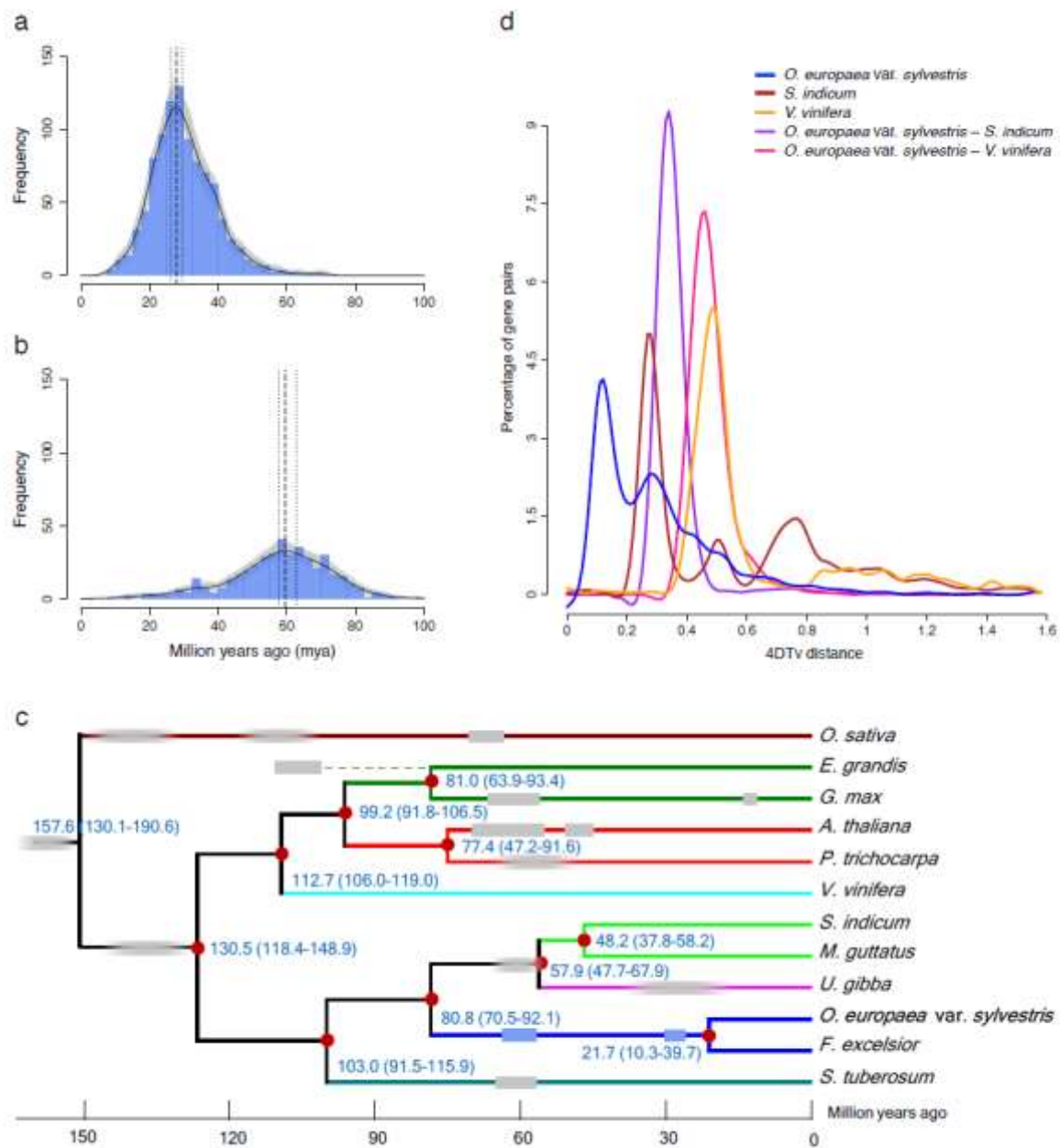


Fig. 2.

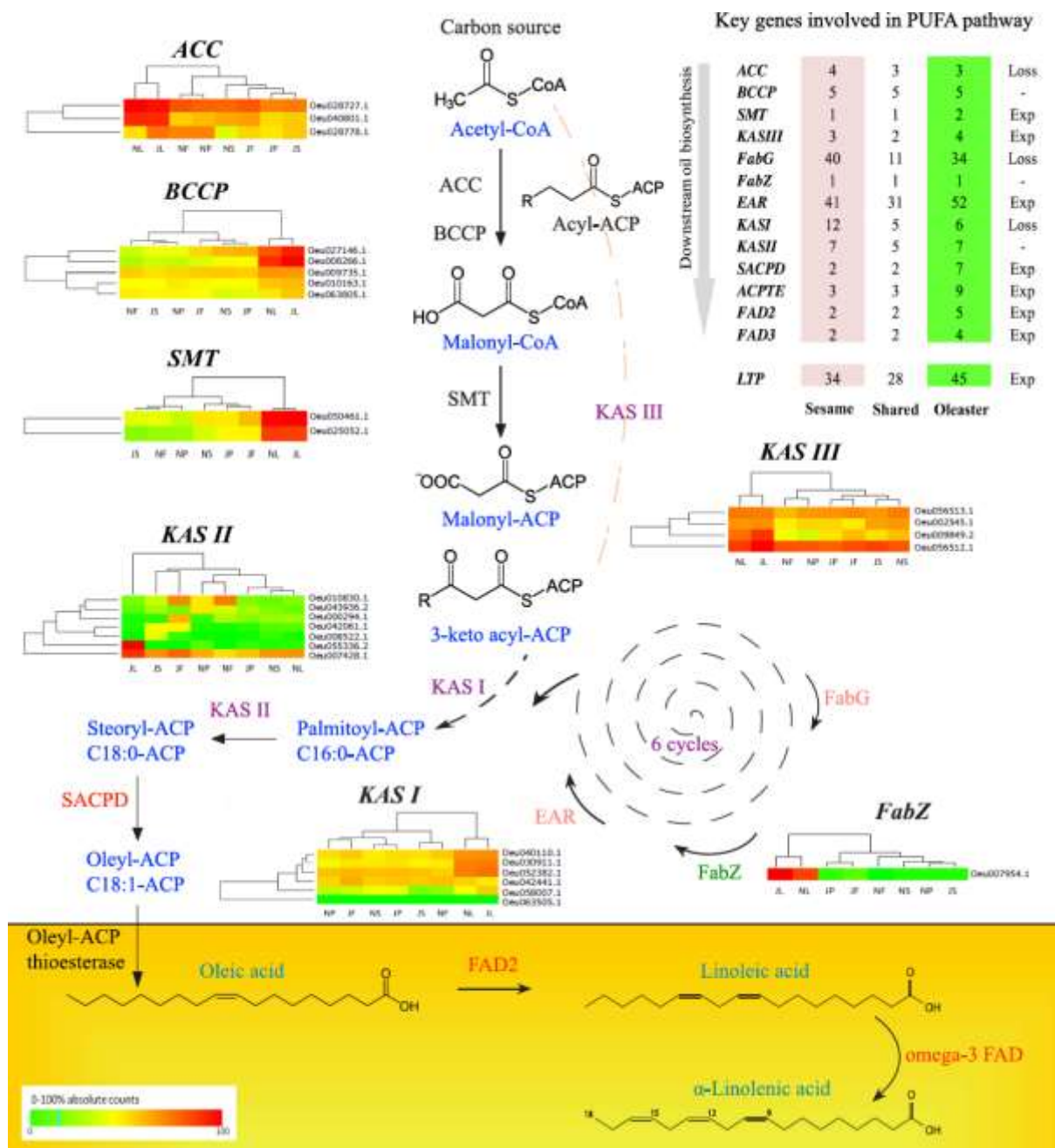


Fig. 3

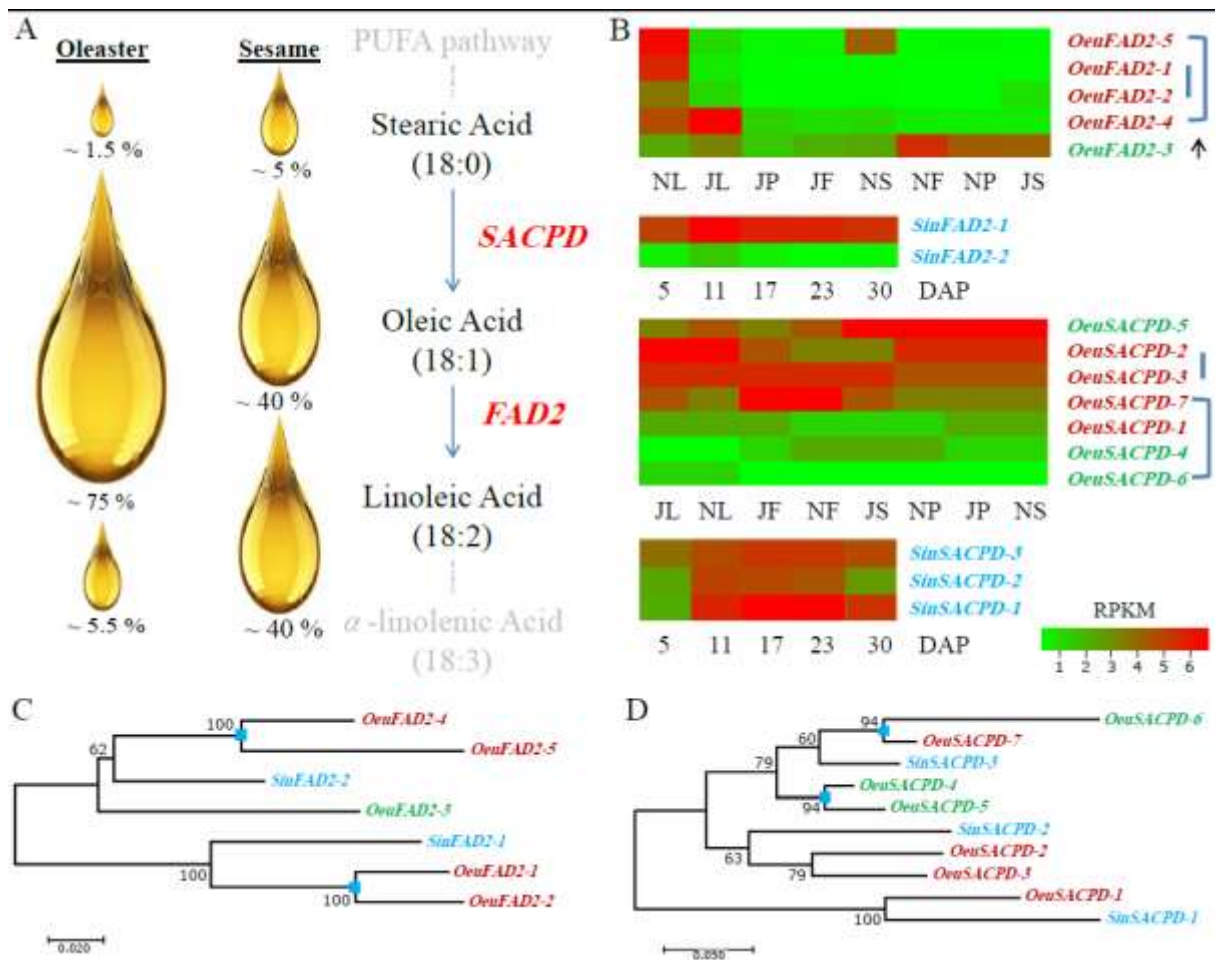


Fig. 4



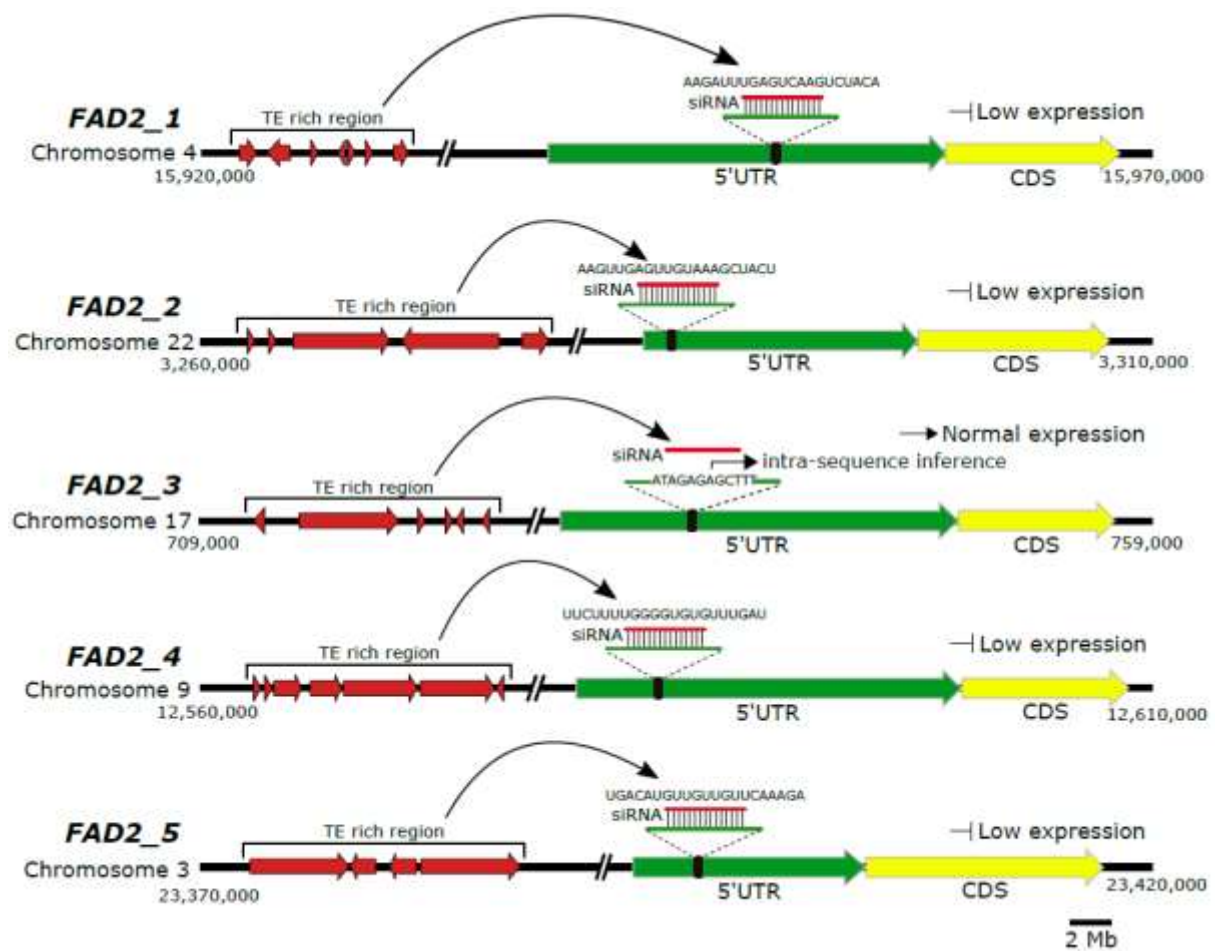


Fig. 5.