SI APPENDIX

S.1. Sample preparation, sequencing and assembly

S.1.1. Sample preparation

Leaves of oleaster (*Olea europaea* L. var. *sylvestris*) (about 2 m high) were collected from the Orhangazi region of Bursa city, Turkey. Leaves were collected and then plants were clonally propagated. All necessary permits were obtained for the described field studies. The DNA was isolated using the modified CTAB method as described in Sahu et al. (2012) (1).

S.1.2. Library construction and genome sequencing

Paired-end (PE) libraries with insert sizes of 250, 500 and 800 bp, as well as 2, 5, 10 and 20 kbp were constructed, following a standard protocol provided by Illumina (San Diego, CA, USA). All these short-reads from short- and long-insert pair-end libraries were subjected for high coverage sequencing on the Illumina HiSeq 2000 platform, resulting in genomic-DNA sequencing data from all 23 libraries (Table S1).

S.1.3. Sequencing data filtering

Prior to the assembly stage, low quality reads were filtered out. These included reads meeting any of the following criteria: (i) >2% ambiguous “N” bases for short insert sizes (250, 500 and 800 bp) or >5% “N” bases for long-insert sizes (2, 5, 10 and 20 kb); (ii) low-quality data for >60% of bases for short insert sizes or >30% of bases for long insert sizes; (iii) >10 bp of adapter sequence; (iv) >10 bp that overlapped between two ends of reads of short insert-size; and (v) identical sequences at both ends. After that, sequenced data totaled up to 220X coverage, given the estimated genome size (Table S1, Figs. S1 and S2).

S.1.4. Error correction

SOAPec v2.01 custom program <http://soap.genomics.org.cn> was used for read-trimming and base correction of 319.19 Gbp clean data from short insert-size libraries.

S.1.5. Genome size estimation of oleaster

S.1.5.1. Flow cytometry

Flow cytometry analyses were performed as recommended by Partec-Sysmex (Kobe, Japan) to estimate the genome size of oleaster (Fig. S1), as described below. Chickpea DNA (*Cicer arietinum* cv. Gokce) was used as a primary internal standard (2C DNA content = 1.65 pg). Briefly, a piece of oleaster leaf and internal-standard leaf were placed in a petri dish with 0.5 ml isolation buffer, sliced with a sharp razor-blade and transferred into a tube. A total of 2 ml of propidium iodide (PI) dye plus RNase were added and the crude nuclear-suspension was filtered through 50 µm nylon mesh. Nuclei were stained with PI using CyStain PI Absolute P. kit. Then, fluorescence intensities of nuclei were measured with a CyFlow SL_3 CD4 counter from the same manufacturer. Filtered suspension was transferred into a new tube and incubated at 37 °C for 30 minutes. Samples were simultaneously measured in the flow cytometer. PI-stained samples were excited at 488 nm with a 15 mW argon-ion laser. Signals of red PI fluorescent area from nuclei were collected in FL2 channel. Mean DNA-content was based on analyses of 10,000 nuclei. G1-peak mean values were used as basis for calculation of absolute DNA amount (2). Histogram of relative nuclear-DNA content was obtained using CyStain PI Absolute P kit. G1 peak of chickpea was detected on channel 132 (740 Mbp for 2C); hence, 1C DNA amount of oleaster was 1.46 Gb.

S.1.5.2. Size estimation by *k*-mer analyses

Genome size was estimated using total length of sequence reads, divided by sequencing depth. The latter was determined by counting the copy number of a certain *k*-mer (e.g., 17-mer) present in sequence reads and plotting the distribution. To calculate the genome size the formula $N \times (L - K + 1)/D = G$ was applied were. N is the total number of sequence-read, L the average length of sequence reads and K is *k*-mer length, (17 bp), G denotes genome size and D the overall depth estimated from the *k*-mer distribution. Additionally, *k*-mer histograms of haploid-sequence data was computed, using the program Jellyfish (3) from Biowire (Mountain View, CA, USA) as follows: jellyfish count -m 17 -o OLEkvmD_k17mer --timing OLEkvmD_k17mer.time -s 4294967296 -t 32 -c 8 -C /dev/fd/0 1>OLEkvmD_k17mer.log 2>OLEkvmD_k17mer.error, jellyfish merge -v -o OLEkvmD_k17mer.jf OLEkvmD_k17mer_* 1>>OLEkvmD_k17mer.log 2>>OLEkvmD_k17mer.error, jellyfish dump -c -t -o OLEkvmD_k17mer.dump OLEkvmD_k17mer.jf 1>>OLEkvmD_k17mer.log 2>>OLEkvmD_k17mer.error, jellyfish stats -o OLEkvmD_k17mer.stats OLEkvmD_k17mer.jf 2>>OLEkvmD_k17mer.error, jellyfish histo -t 32 OLEkvmD_k17mer.jf | sed 's/ / /g' >OLEkvmD_k17mer.histo. Total *k*-mer count was 61,190,425,479; hence, the genome size was estimated (Genome Size = *k*-mer_num/Peak_depth) to be 1.46 Gbp (Table S2).

S.1.5.3. Heterozygosity-rate estimation

The main peak of the 17-mer distribution was around 42X, with an obvious heterozygosis peak near 21X depth. Therefore, a 1.3 heterozygous rate was estimated by *k*-mer distribution of heterozygous sequence (Fig. S2). The highly heterozygous nature of cultivated olive genome (*O. europaea* cv. Farga) situation was also reported as showing a heterozygosis peak near 26X (4).

S.1.6. Genome assembly

All sequence reads were assembled with SOAPdenovo software (5, 6), producing a reference sequence of oleaster genome. A total of 319.39 Gbp of clean data was assembled into contigs and scaffolds, using the de Bruijn graph–based assembler of SOAPdenovo with the following four steps:

i) Building contigs and scaffolds

All reads from short insert-size libraries were used to construct de Bruijn graph with *k*-mer parameter –K37 –R. Then, the graph complexity was simplified by removing tips and connections with low coverage, merging bubbles and masking small repeats. Lastly, *k*-mer path was evaluated to generate the contig file. All usable reads were realigned onto the contig sequences. Amount of shared paired-end relationships between each pair of contigs and rate of consistent and conflicting paired-ends were calculated to construct scaffolds step-by-step, from short insert-size paired-ends to long-insert paired-ends.

ii) Filling gaps

Gaps between contigs were closed by KRSKGF software, version 1.2 (internally developed by BGI) and GapCloser, version 1.10 <http://soap.genomics.org.cn/about.html> for SOAPdenovo. The former is a gap-filling tool based on *k*-mer analyses. GapCloser makes use of reads for local assembly within gaps, by aligning other ends of paired-end reads into scaffolds. Parameters were as follows: GapCloser –a ./O.europaea.scafSeq –b ./O.europaea.lib –o ./

O.europaea.scafSeq.FG –t 64. Scaffolds from reference assembly containing markers genetically mapped in the *SHELL* interval were identified.

iii) Removing redundancy

Rabbit (7) software <ftp://ftp.genomics.org.cn/pub/Plutellaxylostella> was used to remove redundant sequences. It relies on Poisson-based *k*-mer model, which needs a table of *k*-mer frequencies to determine redundant sequences. Jellyfish software was used to generate a table with recommended-parameter K = 17 bp from 250 bp, 500 bp and 800 bp libraries, using the previously described commands (see S.1.5.2 Size estimation by *k*-mer analyses, above). After obtaining the 17-mer occurrence-frequency table, redundancy was removed, using the Redundancy Remover module of Rabbit software.

iv) Reconstructing scaffolds

SSPACE software (8) was used to construct scaffolds. This tool is a stand-alone scaffolder of pre-assembled contigs, using paired-read data. It can use overlapping relationships between contigs and reads to extend contigs, as well as PE relationships of reads to construct scaffolds (Table S3).

S.1.7. Analyses of the genome assembly

S.1.7.1. GC-content distribution

GC content of oleaster and sesame (*Sesamum indicum*) was calculated using 500 bp 250 bp overlapping sliding windows. Main oleaster GC-content peak was around 36.8%, being similar to that of sesame (34.9%), but there was a small curve-shoulder with high GC-content (42% to 44%) (Table 1 and Fig. S3).

S.1.7.2. Analyses of sequencing depth and genome GC-depth

Genome sequencing read-depth distribution was estimated by aligning Illumina reads onto the assembled sequence of oleaster. Mapping was carried out using SOAPaligner <http://soap.genomics.org.cn/soapaligner.html>, with ≤2 mismatches. GC content and average depth with 50 kb non-overlapping sliding windows was calculated. Genome distribution was different from expected pattern, showing a smaller segregated cloud with high GC-content. No obvious difference was found comparing depths of these two blocks (Fig. S4). A total of 417 assembled sequences with 136, 116 and 963 bp lengths, whose GC content ranged from 40% to 50%, were aligned to non-redundant nucleotide (nr/nt) database (20140407) from National Center for Biotechnology Information (NCBI) <http://www.ncbi.nlm.nih.gov>, using megaBLAST 2.2.21 to check for putative spurious-sequences from other species. No obvious contamination from bacteria or fungi was found. The best hits were some repeat sequences, ribosomal DNA (rDNA) encoding ribosomal ribonucleic acid (rRNA), chloroplast and mitochondrion from *O. europaea* and *O. europaea* var. *sylvestris*. Apart from that, no other species were identified, and all alignments were dispersed (Table S4). Distribution of *O. europaea* var. *sylvestris* sequence depth was shown in Fig. S5. Here, filtered reads were aligned onto the genome-sequence assembly using SOAPdenovo2 <http://soap.genomics.org.cn/soapdenovo.html>. Then, the percentage of bases with different depth frequency in genome was calculated. Sequences with coverage under 20X were less than 6%. Average depth was about 180X to 240X. The small peak at 100X to 160X depth may be caused by high levels of genomic heterozygosity (Fig. S5).

S.1.7.3. Assembly evaluation

Transcriptomes from different tissues (leaf, stem, pedicel and fruit, see section S.2.2. further), comprising 212,714 unigenes were mapped with Basic Local-Alignment Search Tool (BLAST) software version 2.2.21 from NCBI using default parameters (9). A total of 91.49% of the unigenes could be aligned to the genome assembly (Table S5a).

S.1.8. Genetic-linkage-map construction and assembly anchoring

Genetic-linkage maps were constructed using the genotyping-by-sequencing (GBS) approach, to develop an integrated genome map for anchoring the assembly (Fig. S6).

S.1.8.1. Genotyping-by-sequencing

DArTseq (10) SBS approach was used to identify single-nucleotide polymorphisms (SNP). In short, 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 after Raman et al. (2014) (11). In short, the following profile was used: one cycle of 94 ºC for 1 min (denaturing); 29 cycles [94 ºC for 20 sec (denaturing, with further ramp of 2.4 ºC/sec to 58 ºC), 58 ºC for 30 sec (annealing, with ramp of 2.4 ºC/sec to 72 ºC) and 72 ºC for 45 sec (polymerization)]; and one cycle of final extension at 72 ºC for 7 min. Samples were then soaked at 10 ºC until removed. PCR, amplicons from each sample were pooled and loaded into cBot (Cluster Station) from Illumina for amplification, using bridge PCR (bPCR). Sequencing of all amplification products was carried out on a single lane of a HiSeq 2000 from the same manufacturer. A single lane generated sequences that were analyzed using proprietary DArT analytical pipelines. In the primary pipeline, FASTQ files were first processed to filter out low-quality sequences.

More-stringent selection, such as $Phred pass score of 30, was applied to barcode region versus rest of sequence. This way, reliable results were obtained during assignments of sequences to specific samples in barcode-split step. Approximately 2,000,000 sequences per barcode/sample were identified and used in marker calling. Finally, identical sequences were collapsed into fastqcall files. These files were used in the secondary pipeline for DArT P/L's proprietary SNP and presence/absence of markers (PAM) calling algorithms (DArTsoftseq). Sequencing data were processed in the analytical pipeline.

S.1.8.2. Linkage-map construction

GBS data were analyzed using regression-mapping algorithm of JoinMap 4.0 software from Kyazma (Wageningen, Netherlands) to enable linkage-map construction. Options for marker placement were determined for each linkage group of two parental-maps using Kosambi's mapping function, minimum logarithm-of-odds (LOD) score threshold of 5.0, recombination-fraction threshold of 0.35, ripple value of 1.0, jump threshold of 3.0 and triplet threshold of 5.0. MapChart 2.0 (12) was used for graphical presentation of linkage maps. In heterozygous perennial trees, such as olives, the strategy used to map F1 populations is the two-way pseudo-test cross-mapping strategy. In this case, marker data analyzed with this strategy and an olive linkage map were constructed separately for each parent (Memecik and Uslu cultivars). 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 cultivar Memecik and cultivar Uslu. For chromosome-scale pseudomolecule construction, two maps were established from two progenies: an F1 progeny of 92 individuals (Memecik × Uslu). An integrated map including 1,307 markers was established (13), based on the double heterozygous loci (14, 15), which were as the benchmarks to judge distances of the marker from different map (Table S5b).

S.1.8.3. Anchoring of genome assembly into genetic map

Several genetic maps, based on either GBS or amplified-fragment length polymorphism (AFLP) markers, were integrated in order to construct linkage groups of the assembled sequences. Genetic markers were mapped onto the scaffolds using Burrows-Wheeler Aligner (BWA) software module for alignment (BWA aln) (16) with default parameters. Afterwards, anchoring of assembled scaffolds to genetic maps was achieved by applying ALLMAPS software (17). All uniquely-mapped markers from either map were provided as input to ALLMAPS with default settings. A total of 1,605 scaffolds were assigned to linkage groups; 516 of them were also oriented (multiple markers on a single scaffold). On the other hand, 41,238 scaffolds (mostly, smaller ones) could not be reliably placed on linkage groups, and were thus categorized as unassigned (Fig. 1a and Fig. S6).

In summary 573 Mbp (~51%) of the current assembly was linked to the genetic maps, based on the unique alignment of 3,491 markers. This generated 23 linkage groups, representing chromosomes in the *O. europaea* var. *sylvestris* genome.

S.2. Genome annotation

S.2.1. Repetitive element annotation

Both homology-based and *de novo* approaches were used to find transposable elements (TE) or transposons, in the oleaster genome. RepeatModeler <http://www.repeatmasker.org/RepeatModeler.html> uses two *ab initio* repeat-prediction programs (RECON and RepeatScout), which identify repeat-element boundaries and family relationships among sequences. PILER (18) and RepeatScout usually work better for *de novo* repeat-library construction of small genomes (≤600 Mbp), while RepeatModeler is more appropriate for larger ones. LTR_FINDER(19) identifies full-length LTR, being used with RepeatModeler to generate the *de novo* repeat library. Tandem repeats were searched for in the genome, using Tandem Repeats Finder (20). TE proteins and *de novo* were combined (Table 1 and Table S6). The homology-based approach involved applying commonly-used databases of known repetitive-sequences, along with programs such as RepeatProteinMask and RepeatMasker (21). The highly repetitive nature of cultivated olive genomes was reported elsewhere (4, 22). According to the *O. europaea* cv. Farga assembly over 63% of the olive genome has repeat elements (4). Consistent with the *O. europaea* cv. Leccino assembly, LTR type TEs are the most common repeat elements (22).

S.2.2. RNA sequencing and assembly of transcriptome data

A total of eight samples were collected from three different individual oleasters located in the central region of Kemalpasa collection orchard (Izmir, Turkey), from leaf, stem, pedicel and fruit tissues in July (2014) and November (2014). Transcriptome sequencing for all oleaster RNA-seq libraries were performed with Illumina RNA-seq protocols. Two assembly approaches were used: *de novo* assembly of clean RNA-reads and reference based using the assembled genome. Gene-expression dynamics were analyzed for development, differentially-expressed genes, expression patterns of positively-selected genes and lineage-specific genes. Paired-end and single-end strand-specific RNA-seq libraries were prepared from these tissues, according to Zhong et al. (2011) (23) and were sequenced on the Illumina HiSeq 2000 system. Raw reads obtained were pre-processed by removing adapter sequences, discarding empty reads and low-quality sequences. All reads were then used for transcriptome *de novo* assembly, using the Trinity short-read assembly program (24). Parameters were "--seqType fq --min_contig_length 100; --min_glue 3 --group_pairs_distance 250; --path_reinforcement_distance 85; and --min_kmer_cov 3". This generated unique sequences of transcripts, often being full-length ones for a dominant isoform, as well as identifying unique portions of alternatively-spliced transcripts. Available RNA-seq data (eight libraries) were aligned to the whole oleaster genome (pseudo-chromosomes and unanchored sequences) using HISAT2 software (25) with default parameters and activated qc-filtering. Resulting compressed binary-versions of sequence alignment/map (SAM) format (BAM) files with alignments were then converted to browser-extensible data (BED) format. Next, coverage tool from BEDTools package (26) was used to calculate the overlap between produced RNA-seq alignments and general-feature format (GFF) file containing gene predictions. A custom Perl <https://www.perl.org> script was then used to calculate coverage of gene-models with different stringency-criteria. The obtained mapping rate of RNA-seq against the genome was between 81% and 85%, depending on the library used. The number of gene models being supported by transcript data varied between 41,559 and 52,819 when no filtering criteria were used (Table S9). When a more stringent criterion was applied (at least 10 reads and >75% of gene model covered), 22,533 to 26,764 gene models were found for a single library and 31,198 (61.5% of predicted) when all RNA-seq libraries were combined. Gene expression levels were calculated by "reads per kilobase transcriptome, per million mapped reads" (RPKM), according to Mortazavi et al. (2008) (27). Comparative gene-expression levels in RNA-seq tissues with false-discovery rate (FDR) ≤0.001 and fold change >2 were evaluated (Fig. S7 and additional data file, see section S.6). To further analyze differentially-expressed genes, Gene Ontology (GO)

<http://geneontology.org> analyses (28) were performed. Lists of such genes across eight different RNA-seq libraries were subjected to GO enrichment, by mapping all listed genes to GO database. Then, calculating gene numbers for every term with hypergeometric test, significantly-enriched GO terms were identified, based on GO-TermFinder <http://go.princeton.edu/cgi-bin/GOTermFinder>. After that, the following three ontologies were identified: molecular function, cellular component and biological process (additional data file, see section S.6). A total of 8,469, 15,102, 8,848 and 10,215 genes were regulated in fruit, leaf, pedicel and stem tissues, respectively (additional data file, see section S.6). GO-enrichment analyses were performed for differentially-expressed genes in ripe fruit tissues collected in November (ripe or mature), compared to unripe-fruit tissues collected in July (unripe or immature). Differences were mostly related to oil-producing fruit tissue, including terpenoid and isoprenoid metabolism, lipid metabolism and transport processes. As expected, this further confirms the oil and secondary-metabolite biosynthesis and transport in mature fruit (additional data file, see section S.6). On the other hand, catalytic, ion binding and transferase-activity genes in pedicel tissue were preferentially expressed. Interestingly, stem tissue gene-expression GO terms were mostly related with response stimuli, as well as hormone elicitation. Fatty-acid biosynthesis and transport processes were also identified in stem tissue gene-expression patterns, which were differentiated for oil biosynthesis (additional data file, see section S.6).

S.2.3. Gene annotation

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 (29) was used to consolidate results. Protein sequences of *Arabidopsis thaliana, Sesamum indicum* (sesame), *Solanum tuberosum* (potato) and *Vitis vinifera* (common grape vine) were aligned with TBLASTN and genBLASTA (30) against the matching proteins using GeneWise (31) for accurate spliced alignments. Next, we used the *de novo* gene-prediction methods GlimmerHMM (32) <https://ccb.jhu.edu/software/glimmerhmm> and Augustus (33) to predict protein-coding genes, using parameters trained for *O. europaea* var. *sylvestris, A. thaliana, S. indicum*, *S. tuberosum* and *V. vinifera* (Table S8).

S.2.4. Non-coding RNA annotation

The assembled oleaster genome was screened for non-coding RNA (ncRNA). Thus, tRNA genes were predicted by tRNAscan-SE version 1.23 (34) with eukaryote parameters. On the other hand, rRNA template sequences from plant-rRNA databases were aligned against the *Olea europaea* var. *sylvestris* genome using BLASTN, to identify putative rRNA. Other noncoding RNA, like microRNA (miRNA) and small nuclear-RNA (snRNA), were identified using INFERNAL version 0.81 (35) by searching against the RNA family (Rfam) (36) database release 10.1 <ftp://ftp.sanger.ac.uk/pub/databases/Rfam> (Table S10).

S.2.4.1. miRNA identification

Raw sequences from six small RNA (sRNA) libraries generated by our group (37) were used for miRNA identification. sRNA reads were first cleaned by filtering low-quality ones and removing adapters. A restrictive approach was carried out to avoid false positives: only sequences present in at least four out of six sequencing libraries were kept and combined together for subsequent analyses. miRDeepFinder (38) was used to perform sRNA analyses (including categorizing conserved-miRNA, identifying miRNA and their targets), as well as GO/Kyoto Encyclopedia of Genes and Genomes

(KEGG) (39) <http://www.genome.jp/kegg> for annotation of miRNA and their targets. Briefly, other small RNA, including rRNA, snRNA and tRNA, were discarded if they could be fully mapped to Rfam 10.1. To differentiate conserved and non-conserved reads, remaining reads were aligned against all known plant miRNA (miRBase, release 21) (40) by Water module of European Molecular Biology Open Software Suite (EMBOSS) package (41). A conserved read is defined to have no more than three mismatches with known miRNA sequences; otherwise being considered a non-conserved read (42). A putative miRNA star [miRNA*; also known as minor miRNA, corresponding to the other (antisense) strand of the major miRNA] sequence whose 3'-OH end has two-nucleotide (nt) overhang is considered a real one. However, it should be taken into account that some well-known miRNA do not always co-exist with their miRNA* in small sequencing-libraries. Each of six sRNA sequencing libraries generated ~15 million reads on average, representing ~6 million unique-sequences (Table S11). A total of 11.21% non-redundant sequences had more than three reads, accounting for 57.87% in total. Of these sequenced (≥3 reads), tRNA, rRNA and snRNA had average ratios of ~ 0.94%, 2.78% and 0.21% of total reads, respectively, amongst the six libraries, while miRNA accounted for an average of 15.06% of total reads and 0.21% of unique reads (Table S11). An averaged total of ~83.40% (≥3 reads) could be completely mapped back to oleaster genome, corresponding to 75.51% unique reads on average.

Most reads in the six libraries had a length between 18 and 26 nt. The most abundant ones were of length 24 nt, followed by 21 nt. Furthermore, nucleotide-base distributions in 18 to 26 nt reads were found to be similar across the six libraries. A/GC/U content of reads with length of 22 and 23 nt in the July-Fruit (JF) library were higher than those in the other five. Likewise, A/GC/U content of reads with length of

21 nt in the November-Leaf 1 (NL1) were higher than those in the other five libraries (Fig. S8). We have previously found that mature olive-tree miRNA preferentially start with U (37). That was confirmed in the present work with oleaster, especially for reads with length of 20 and 21 nt. On the other hand, C was the dominant base at the end of small RNA in the six libraries, particularly for 20 and 21 nt reads. Thus, the start and end nucleotide of small RNA in this work are likely due to miRNA.

A total of 498 conserved-miRNA families were identified in the six wild-olive-tree libraries, including miR156, miR157, miR160, miR166, miR168, miR172, miR394 and miR399. In general, conserved miRNA in July-Leaf 1 (JL1) versus NL1 and JF versus NL2 showed similar expression (Fig. S9). On the other hand, miR156, miR157, miR164, miR166, miR167, miR168, miR172 and miR391 consist of top high-expression miRNA families. Pearson's chi-squared test detected 334 (67.8%) conserved-miRNA families differentially-expressed in the six oleaster libraries (*p*-value ≤0.01). Compared with other species from both eudicotyledons and monocotyledons, oleaster has relatively broad conserved-miRNA families, including miR156, miR157, miR158, miR159, miR160 and miR163 (Fig. S10). A total of 213 conserved wild-olive-tree miRNA precursors were identified from 203 contigs, belonging to 94 families. Only 24 miRNA had their miRNA*, with 18 (69), 21 (63) and 19 nt (44) miRNA being the largest ones. The length of conserved precursors varied from 50 to 888 nt. A miRNA star was required to co-exist at least in one sequencing library. Thus, 125 novel miRNA were identified in oleaster from the six libraries. Most of such novel miRNA only had one member, excluding miR-N5, miR-N8, miR-N9, miR-N14, miR-N15, miR-N21, miR-N24, miR-N25, miR-N35, miR-N49, miR-N57, miR-N60 and miR-N99.

S.2.4.2. miRNA-target transcript analyses

In this study, only novel miRNA and their miRNA* were required to co-exist in at least one sequencing library. psRNATarget <http://plantgrn.noble.org/psRNATarget> (43) was used to identify miRNA targets with default parameters. Next, the miRNA targets were subjected to GO term classification. KEGG-pathway enrichment was performed following KEGG annotation attached in GO-protein annotation. Pearson's chi-squared test was used to differentiate miRNA expression in the six sequencing libraries. Considering highly conserved miRNA and their functions, all conserved miRNA (3,322 unique reads) and 125 novel ones were used to predict miRNA targets by psRNATarget, resulting in 29,842 miRNA-target pairs including 7,849 unique target-genes in total. Of these targets, 277 miRNA families targeted 7,322 genes and 778 genes were targeted by 108 novel miRNA. According to miRNA major function type, these miRNA targets were categorized into six function groups, including transcription factor (TF), development, hormone, metabolism, signal transduction and stress response (Table S12). It turned out that 4,606, 1,937 and 630 miRNA targets were associated with transcription factor, stress response and metabolism, respectively. A total of 359 miRNA and their 4,413 targets were categorized into 302 cellular components, 1,498 biological processes and 898 molecular functions. Further KEGG pathway analyses showed that 1,101 miRNA-target pairs were enriched in 175 pathways.

S.2.5. Functional annotation

Gene-function information, protein motifs and domains were assigned by comparing with public databases, including Swiss protein (SwissProt) <http://www.expasy.ch/sprot> sequence database and its supplement TrEMBL (44), KEGG, InterPro (45) and GO (28). Target locations of homolog proteins were obtained by aligning protein sequences of *A. thaliana*, *S. indicum*, *S. tuberosum* and *V. vinifera* to the *O. europaea* var. *sylvestris* genome, using TBLASTN with expect (E)-value parameter of 1×10^{-5} (Table S13). Genes were divided into two classes for gene-family clustering: i) one is cluster, which is represented with CL prefix following with cluster identity (ID). In a single cluster, there are several unigenes showing high similarity (>70%); ii) The other one is singleton, which is represented as unigene. Assembled unigenes were aligned to protein databases, such as NCBI nr protein, Swiss-Prot protein, KEGG pathway and Clusters of Ortholog Groups (COG) <https://www.ncbi.nlm.nih.gov/COG> by translated BLAST alignment with expected E-value < 10^{-5}, searching protein databases using a translated nucleotide query (BLASTX). The best aligning results were used to decide sequence direction of unigenes. A priority order of nr, Swiss-Prot, KEGG and COG was considered to resolve any contradictory result between different databases. On the other hand, ESTScan program <http://www.ch.embnet.org/software/ESTScan.html> was used to predict coding sequences (CDS) and orientation of sequences that failed to be annotated to any database. Functional annotations of unigenes by GO were carried out using BLAST2GO software (46), followed by GO functional classifications using the WEGO software (47) (additional data file, see section S.6). That allowed to view the distribution of gene functions in *O. europaea* var. *sylvestris* at the macro level and to functionally compare annotations of oleaster with ten plant species (*A. thaliana*, *S. indicum*, *S. tuberosum*, *V. vinifera, Populus trichocarpa* (poplar), *Prunus persica* (peach), *Oryza sativa* (rice), *Eucalyptus grandis* (flooded gum)*, Mimulus guttatus* (monkey flower) and *Glycine max* (soybean)) downloaded from Phytozome (48) v.9 <https://phytozome.jgi.doe.gov/pz/portal.html> and *Utricularia gibba* (bladderwort) taken from <https://genomevolution.org/CoGe/OrganismView.pl?oid=36222> and

Fraxinus excelsior (ash tree) (49) was downloaded from <https://www.ncbi.nlm.nih.gov/genome/?term=fraxinus+excelsior>. Three GO ontologies (cellular component, biological process and molecular function) were analyzed at level 5 using Gypsy Database Professional (GPRO) 1.0 (50). All genes were mapped to terms in KEGG (51) database using default parameters. Gene Ontology (GO) analysis resulted in the successful annotation of 72.42% of all genes with 1,516 enzyme-encoding mappings from 141 KEGG pathways. Binding (23,611) was the most represented molecular function, and 21,870 genes were annotated for metabolic processes in the oleaster genome. Functional comparison of GO-based annotations among oleaster and eight other sequenced genomes showed those two functions also as the most represented in *G. max* and *P. trichocarpa*.

S.3. Evolutionary analyses

S.3.1. Ortholog gene clusters

Genes of 12 sequenced species (*O. europaea* var. *sylvestris, A. thaliana, E. grandis, G. max, O. sativa, P. trichocarpa, S. indicum, S. tuberosum, V. vinifera, M. guttatus, F. excelsior and U. gibba*) were used for gene family clustering analysis. Gene families of orthologous genes were determined with the OrthoMCL software (52) with default settings except for the inflation factor which was set at 3. The input for OrthoMCL was the result of an all-versus-all BLASTP analysis of the protein sequences of 11 selected plant species with the proteins of oleaster (*O. europaea* var. *sylvestris*, this study) added as the 12th species, resulting in 436,088 proteins in total. BLASTP was run with an E-value cutoff of 1 × 10^{-5} and with number of reported alignments set at 10,000. Protein clustering of the predicted olive genes with the 11 other species resulted in 17,208 multi-species gene families for olive, with an additional 1,070 olive-specific gene families. The number of gene families is largely consistent over the different species and 8,986 oleaster genes were not assigned to any gene family and remained unclustered (Fig. S11, Table S14a and S14b).

S.3.2. Phylogenetic analyses

A phylogenetic tree was constructed for evolutionary analyses, on the basis of a concatenated sequence-alignment of 231 single-copy genes shared by oleaster and 11 other plant species (*O. sativa, A. thaliana, P. trichocarpa, G. max, E. grandis, S. tuberosum, V. vinifera, S. indicum, F. excelsior, M. guttatus and U. gibba*). Multiple alignments of protein sequences was carried out with MUSCLE (53) for each single-copy gene family, further converting protein alignments into coding sequences using a Perl script. Phase-1 sites were extracted from each family and concatenated to one supergene for every species, and MrBayes 3.1.2 (54) was used to construct the phylogenetic tree based on the GTR+gamma evolutionary model (Fig. 2c). In the resulting phylogenetic tree, *U. gibba* diverged before *S. indicum*. This placement is inconsistent with the consensus APG IV phylogeny (55) in which Pedaliaceae (*S. indicum*) diverged first, followed by the divergence of Lentibulariaceae (*U. gibba*) and Phrymaceae (*M. guttatus*) (56, 57). A high substitution rate in *U. gibba* (58) might be the reason for this potentially erroneous placement, see also Sollars et al. (2017) (49) and He et al. (2016) (59).

S.3.3. Estimation of divergence time

The constructed phylogenetic tree was used to calculate divergence time among species. The calibration tree was generated using a fossil date for the split of *A. thaliana* and *V. vinifera* from TimeTree <http://timetree.org>, calibrating gene-evolution rate in dendrogram, using penalized likelihood method with truncated Newton-optimization, as implemented by r8s version 1.71 (60). Then, divergence time of each tree node was inferred using Bayesian Markov-chain Monte Carlo (MCMC) tree (MCMCTree) package in PAML (61) with the JC69 model.

S.3.4. Whole genome duplications

Duplications of oleaster, *S. indicum* and *V. vinifera*, and speciation events between oleaster and either sesame or grape vine, were analyzed via 4DTv approach. Concatenated nucleotide alignments were calculated with HKY85 substitution models (Fig. 2d). Synonymous substitutions per synonymous site (K_S)-based age distributions of oleaster were also constructed to estimate duplication events in the oleaster genome, as previously described by Vanneste et al. (2013) (62). Briefly, the paranome (set of all duplicate genes belonging to gene families in a genome) was constructed by performing an all-against-all protein-sequence similarity search using BLASTP with an E-value cutoff of 1×10^{-10}. Then, gene families were built using the mclblastline pipeline version 10-201 <http://micans.org/mcl> (63). Gene families of size larger than 100 were excluded. Each gene family was aligned using MUSCLE version 3.8.31. K_S estimates for all pairwise comparisons within a gene family were obtained through maximum likelihood (ML) estimation, using CODEML (64) from the PAML package version 4.4c (65). Gene families were then subdivided into subfamilies for which K_S estimates between members did not exceed a value of 5. To correct for the redundancy of K_S values (a gene family of n members produces $n(n–1)/2$ pairwise K_S estimates for $n–1$ retained duplication events), a phylogenetic tree was constructed for each subfamily using PhyML (66) under default settings. For each duplication node in the resulting phylogenetic tree, all m K_S estimates between the two child clades were added to the K_S distribution with a weight of $1/m$ (where m is the number of K_S estimates for a duplication event), so that the weights of all K_S estimates for a single duplication event sum up to one. The resulting age distribution of the oleaster paranome is shown in Fig. S12a. Paralogous gene pairs located in duplicated segments (anchors), assumed to be corresponding to the most recent WGD(s), were detected using i-ADHoRe version 3.0 (67, 68). The distribution of K_S values between such anchor-pair genes is shown in Fig. S12b. Identified anchor pairs confirmed the presence of two WGD peaks around a *K*S of 0.25 and 0.75, respectively (the long tail and additional small humps in the anchor-pair distribution are most likely due to small saturation effects (62) and the remnants of older WGDs in the eudicot lineage, such as the shared pan-eudicot gamma triplication event).

Absolute dating of the two identified WGD events in oleaster was performed as previously described by Vanneste et al. (2014) (69). Briefly, two kinds of duplicated gene pairs were collected for phylogenetic dating: i) paralogs from duplicated segments (anchor pairs, as identified by i-ADHoRe); and ii) non-anchor paralogs, lying under the WGD peaks identified in the K_S distributions (peak-based duplicates). The two WGDs were dated separately: anchor pairs and peak-based duplicates with K_S values between either 0.15 and 0.45 (most recent WGD) or 0.55 and 0.95 (older WGD) were selected for absolute dating (dashed lines in Figs. S12a and S12b). For each WGD paralog pair, an orthogroup was created including both paralogs plus several orthologs from other plant species, as identified by InParanoid (v4.1) (70). A broad taxonomic sampling was used: a single representative ortholog from the orders Cucurbitales, Malvales and Solanales, and two representative orthologs from the orders Rosales, Fabales, Malpighiales, Brassicales and Poales. A total of 90 and 360 orthogroups based on anchor pairs and peak-based duplicates, respectively, were collected for the older WGD.

On the other hand, a total of 1,000 orthogroups based on anchor pairs were collected for the younger WGD. The node joining any two *O. europaea* var. *sylvestris* WGD paralogs was then dated using Bayesian Evolutionary Analysis Sampling Trees (BEAST version 1.7 package) (71) under an uncorrelated relaxed clock model and a LG+G (four-rate categories) evolutionary model. A starting tree with branch lengths satisfying all fossil prior constraints was created according to the consensus APG IV phylogeny (55). The same fossil calibrations were used as for dating the asterid WGDs in Vanneste et al. 2014 (69). A run without data was performed to ensure proper placement of the marginal calibration prior (72). The Markov chain Monte Carlo (MCMC) for each orthogroup was run for 10 million generations, sampling every 1,000 generations, resulting in a sample size of 10,000. The resulting trace files of all orthogroups were evaluated automatically using LogAnalyser (part of the BEAST package (71)) with a burn-in of 1,000 samples to ensure proper convergence (minimum ESS for all statistics at least 200). A total of 943 orthogroups were accepted for the most recent WGD and the absolute age-distribution of the estimates for the node uniting the WGD anchor pairs is shown in Fig. 2a. On the other hand, too few anchor pairs were available to evaluate them separately from the peak-based duplicates for the older WGD. Thus, absolute age-estimates from the orthogroups based on both anchor pairs and peak-based duplicates were grouped into one absolute age distribution (Fig. 2b); in total, 426 orthogroups were accepted. Kernel-density estimation (KDE) and a bootstrapping procedure were used to find peak WGD age-estimates and their 90% confidence-interval boundaries, respectively. More detailed methods are available in Vanneste et al. (2014) (69).

To test whether the older WGD is either an independent event specific to the Oleaceae lineage or represents a shared event in the common ancestor of oleaster and

S. indicum we built and examined gene trees containing paralogous genes duplicated in this event. We used OrthoFinder version 0.7.1 (73) to identify orthologous groups with proteins from oleaster, *S. indicum*, *S. lycopersicum*, *S. tuberosum* and *V. vinifera*. Multiple sequence alignment for each gene family were built using MUSCLE version 3.8.31, based on amino acid sequences, trimmed with a heuristic mode (-automated1), and backtranslated into a nucleotide sequence alignment with trimAl version 1.4 (74). Maximum likelihood (ML) gene trees were then inferred by RAxML version 8.2 (75) using the GTR+GAMMA+I model with 100 rapid bootstrap analysis. From these gene trees, we selected the ones that contained paralogous gene pairs from duplicated, collinear regions (anchor pairs) with K_S values in the range of 0.55 to 0.95 (i.e., duplicates from the older WGD peak in the K_S distribution, Fig. S12b). Gene trees were further required to include genes from oleaster, *S. indicum*, *V. vinifera* and at least one gene from *S. lycopersicum* or *S. tuberosum*. In total, we obtained 593 such gene trees containing 751 anchor pairs. We used *V. vinifera* genes to root the gene trees if they formed a monophyletic group, otherwise, gene trees were rooted by its mid-point. Each gene tree was then searched for (sub)topologies with duplication events that were supported by oleaster anchor pair(s) and where the sister lineage(s) according to the species phylogeny (based on Fig. 2c) were consistent with either one of the two WGD scenarios. For example, a duplication before the divergence of oleaster and *S. indicum* but after the split of Solanaceae that is supported by oleaster anchor pair(s) would be considered as a topology in support of a WGD shared by oleaster and *S. indicum* (Figs. S13b and S13c). Bootstrap values of the identified duplication event(s) and the oleaster–*S. indicum* speciation event(s) were used to evaluate confidence in a particular duplication event. For instance, in the topology (((Oeur1,Oeur2)*bootstrap1*,(Sind1,Sind2)*bootstrap2*)*bootstrap3*), *bootstrap1* is the bootstrap value supporting the oleaster duplication and *bootstrap3* is the bootstrap value supporting the oleaster–*S. indicum* speciation. Such a topology would be considered as providing support for an independent WGD in the Oleaceae lineage if *bootstrap1* and *bootstrap3* are greater than or equal to 50%. Similarly, in the topology (((Oeur1,Sind1)*bootstrap1*,(Oeur2,Sind2)*bootstrap2*)*bootstrap3*), *bootstrap3* is the bootstrap value supporting a duplication event shared by oleaster and *S. indicum*, and *bootstrap1* and *bootstrap2* are the bootstrap values supporting the subsequent speciation event. If *bootstrap3* and at least one of *bootstrap1* and *bootstrap2* were greater than or equal to 50%, the topology would be considered as providing support for a shared WGD between oleaster and *S. indicum*.

The two scenarios for the older WGD were supported by very similar numbers of gene trees (Table S15, rows 1 and 2). Therefore, we used two additional approaches to check the set of gene trees from each scenario for consistency. First, we evaluated the identified (sub)topologies under both scenarios for additional support from *S. indicum* anchor pair(s) (as detected by i-ADHoRe version 3.0 (67, 68)). Sub(topologies) in which both *S. indicum* anchor pair(s) correspond to an independent duplication event specific to *S. indicum* and oleaster anchor pair(s) correspond to an independent duplication event specific to the Oleaceae lineage were considered consistent (e.g., ML gene tree in Fig. S13a). Sub(topologies) in which oleaster anchor pair(s) correspond to a WGD shared by oleaster and *S. indicum* but *S. indicum* anchor pair(s) correspond to an independent duplication event specific to *S. indicum* (e.g., ML gene tree in Fig. S13b) were considered inconsistent as there is no evidence for two WGDs in the *S. indicum* lineage. Interestingly, most of the *S. indicum* anchor pair(s) indicated duplications specific to *S. indicum* in trees for both scenarios, and based on this 95 of the 185 (sub)topologies supporting a shared WGD were considered inconsistent (Table S15, row 3). Second, we used the Approximately Unbiased (AU) test (76) to further assess the confidence of the gene trees. The ML topologies were simplified into testable trees that contained only an anchor pair from oleaster, a single ortholog each from *S. indicum* and *V. vinifera*, and a single ortholog from either *S. lycopersicum* or *S. tuberosum*. Alternative trees of these simplified topologies were generated based on the rival WGD scenario. For a simplified ML topology that supports an independent Oleaceae WGD, two alternative trees exist in which the sesame gene clusters with either one of the anchor pair genes of oleaster (Fig. S13a). Only one alternative tree exists for a simplified ML topology that supports a shared WGD in the common ancestor of oleaster and *S. indicum* (Figs. S13b and c). All sets of simplified and alternative trees were then tested using the AU test as implemented in CONSEL version 0.20 (77). Site-wise log-likelihoods for the trees were calculated using RAxML under the GTR+GAMMA+I model. A ML topology was considered exclusive (i.e., it has high confidence) if all the alternative trees were significantly rejected by the AU test (i.e., all have *P*-values < 0.05, see for example Figs. S13a and b).

S.3.5. Syntenic analyses and whole-genome alignment

A BLASTP search (with an E-value cutoff of 1×10^{-5}) was performed to identify paralogous genes. Syntenic blocks (with at least five genes per block) were identified by MCscan (78) with parameters of "-a -e 1e-5 -u 1 -s 5". For the alignment results between these, each aligned block represented the ortholog pair derived from the common ancestor. Sequences that contained genes were used to show intergenomic relationships with their length information. Fourfold-degenerate values of blocks were calculated as revised by HKY85 model. Whole-genome alignment was carried out by LASTZ (see URL <http://www.bx.psu.edu/~rsharris/lastz>) between *S. indicum* and

O. europaea var. *sylvestris*, after repeat regions were masked. SyMAP (79, 80) was used to compute the synteny blocks between the wild olive genome and the genomes of other species (i.e., *S. indicum, V. vinifera, P. trichocarpa and S. tuberosum*). Sequences of each genome were aligned and the raw anchors are computed using MuMmer (81). MuMmer was used with the PROmer operation mode. Raw hits arisen from the large number of repetitive sequences were reduced by clustering the resulting anchors into gene anchors, further applying the reciprocal-top-2 filtering method. Circos v0.68 (82) was used to produce circular visualization of the oleaster genome features. Custom Perl scripts were used to convert the assembly outputs into the Circos format. A total of 37,789 and 36,677 anchors (orthologous hits) were identified comparing the sesame and the grape genomes to oleaster, respectively. These anchors spanned 5% and 3% of the oleaster genome. A total of 917 and 543 syntenic blocks were detected between oleaster and poplar and potato, respectively (Figs. S14 and S15, and Table S16).

S.3.6. LTR-insertion date estimations

S.3.6.1. Identification of full-length LTR retrotransposons

Assembled contigs were searched *de novo* for identification of full-length long-terminal-repeat retrotransposons. For this purpose, oleaster sequences were analyzed using the LTR_FINDER software (83). Alignment boundaries of ends of LTR-pair candidates were adjusted using the Smith–Waterman algorithm. Boundaries were readjusted based on occurrence of the following criteria: i) being flanked by TG and CA dinucleotides at 5’- and 3’-end, respectively; ii) presence of target-site duplication (TSD) of 4 to 6 nt; iii) putative 15- to 18-nt primer binding-site (PBS), complementary to tRNA at the end of putative 5’-LTR; and iv) 20- to 25-nt polypurine tract (PPT) just upstream of 5’-end of 3’-LTR by CAP3 software (84) under relaxed settings (-o 30, -p 80, -s 500), in order to reduce redundancy. Identified sequences were searched against public non-redundant databases at NCBI and Repbase (85) using BLASTN and BLASTX with E-value thresholds of $<10^{-5}$ and $<10^{-10}$, respectively.

S.3.6.2. Estimation of insertion time of full-length LTR retrotransposons

Since both LTR sequences bordering LTR-retrotransposons are identical at the time of insertion, the 5’- and 3’-LTRs of each putative retrotransposon can be compared to estimate the insertion time of full-length LTR retrotransposons. The synonymous substitution-rate was calculated by comparing single-copy genes of *S. indicum* to oleaster ortholog genes. As the estimated separation between *O. europaea* var. *sylvestris* and *S. indicum* was dated to 80.8 (70.5–91.2) mya (Fig. 2c), it was used to estimate the synonymous nucleotide-substitution rate applying the “T = K/2r” formula where r is the number of nucleotide substitutions per site, per year; K is the number of substitutions per site between two homologous sequences; and T is the time of divergence between two sequences. Given the divergence-time estimation between oleaster and sesame, T = 80.8 Mya, and K = 0.188; thus, $r = 1.32 \times 10^{-3}$. The two LTRs of each full-length retrotransposable element (RE) were aligned and indels were eliminated. Nucleotide-substitution rates were calculated using MEGA6 software (86). Insertion time for each full-length retrotransposon was estimated as proposed by Ma and Bennetzen (2004) (87) using twice the mean number of synonymous substitutions (per site, per year) as nucleotide-substitution rate between LTRs (Fig. S16).

S.3.7. Ortholog and in-paralog genes between oleaster and sesame

InParanoid program 4.1 (88) was used to identify common groups of ortholog and in-paralog genes involved in oil biosynthesis pathways, between oleaster and sesame. First, a list of oleaster oil-biosynthesis genes was created. Thus, lists of gene names related to lipid (oil/fat) biosynthesis were obtained from scientific literature and protein databases available at NCBI. Such sequences were used as query via BLASTX (89) against Reference Sequence (RefSeq) Database (release 77) <ftp://ftp.ncbi.nlm.nih.gov/refseq/release> (90). The BLAST pipeline implemented in GPRO 1.0 (91) was used in order to obtain comprehensive information about presence of encoding genes in oleaster, as well as number of functional loci. An E-value threshold of 10^{-10} was used to filter false positives. Then, interologs were determined using the InParanoid ortholog-predicting algorithm, by means of pairwise similarity scores. They were calculated with BLAST between oil-biosynthesis gene dataset of *O. europaea* var. *sylvestris* and *S. indicum* proteome (92) datasets. Results were generated in Fast Alignment Sequence Tools (FAST)-All (FASTA) format for constructing orthology groups. Such predicting algorithm separated orthologs from out-paralogs (homologs resulting from a duplication event predating speciation between two species). Ortholog groups were built with InParanoid, setting the two seed-orthologs first, by two-way best hits between two datasets. Then, sequences that were closer to corresponding seed-orthologs than to any other sequence in other dataset were added. The program provides a confidence value for each in-paralog, showing how closely related it is to its seed ortholog. In addition to confidence values for in-paralogs, reliability of each ortholog group was estimated. Thus, the bootstrapping technique was used with the following parameters: score cut-off as 40 bits, in-paralogs with confidence less than 0.05 not shown, sequence overlap cut-off as 0.5, group merging cut-off as 0.5 and BLOSUM62 matrix. The gene dataset was functionally annotated, using BLAST combined to protein-domain approaches (93), allowing to validate results in the absence of any functional information about the *S. indicum* proteome, as well as to analyze the shared and unique genes involved in fatty-acid biosynthesis.

Protein sequences from the *O. europaea* var. *sylvestris* oil-biosynthesis dataset were compared to those in the proteome of *S. indicum* in a pairwise fashion, using reciprocal BLAST to separate in-paralogs from orthologs and out-paralogs. A total of 2,025 genes, out of 2,328 oleaster oil biosynthesis genes had homologs in the sesame genome. A total of 911 ortholog groups were built (after excluding out-paralogs), including 1,232 in-paralogs, with multiple possible-orthologs and strict one-to-one orthology from *O. europaea* var. *sylvestris*, as well as 1,171 in-paralogs from *S. indicum*. The 563 oil-biosynthesis genes conserved in a strict one-to-one orthology between oleaster and sesame did not undergo any additional duplication events. The rest of in-paralogs (669 and 608, respectively) were the result of duplication events, following speciation. Such results support the suggested genomic evolution and duplication events in such species (Figs. 2a-d).

Olive oil is mainly composed of triacylglycerols (TAG), which contain fatty-acid residues after esterification with glycerol (94). Monounsaturated oleic acid (C18:1) represents ~75% in olive oil, followed by saturated palmitic acid (C16) (~13.5%), polyunsaturated linoleic acid (c18:2 ω-6) (~5.5%) and α-linolenic acid (c18:3 ω-3) (~0.75%) (94). Oil biosynthesis consists of three major steps; fatty-acid biosynthesis, acyl editing and triacylglycerol (TAG) synthesis (95). Fatty acids are synthesized starting from a photosynthate long-chain, further modified and degraded by enzymes encoded by a large number of genes, including fatty acid synthases, elongases, desaturases and carboxylases (96). A number of genes involved in lipid metabolism in plants have been reported, including oils and fats (liquid or solid at room temperature, respectively) (96). Fatty-acid biosynthesis is one of the main steps of the complex lipid/oil metabolism (95), including elongation, degradation, biosynthesis of unsaturated fatty acids and linoleic-acid metabolism.

Since one of the main steps of the complex TAG-generation in oleaster is fatty-acid biosynthesis, genes involved in the latter (including unsaturated fatty acids and linoleic-acid), elongation and degradation were annotated in both *O. europaea* var. *sylvestris* and *S. indicum* genomes. KEGG metabolic-annotation based on oil biosynthesis genes of oleaster (2,328) and sesame (1,161) resulted in 1,267 and 273 matching fatty-acid biosynthesis-related pathways, respectively. Sequence hits of dehydrogenases, monooxygenases and oxidases were highly represented in sesame. The major differences in gene-count derived from sequences encoding dehydrogenases (727 in sesame and 11 in oleaster) for fatty-acid degradation, as well as the ones encoding monooxygenases (96 in sesame and four in oleaster) for linoleic-acid metabolism. Additionally, InParanoid allowed identifying unique and shared genes involved in these pathways. For example, 14 of 26 genes were found to be unique to oleaster for biosynthesis of unsaturated fatty acids. On the other hand, *O. europaea* var. *sylvestris* had more key genes involved in oil biosynthesis than *S. indicum*, except for dehydrogenases and isomerases (Fig. S17).

S.3.8. Phylogenetic analyses based on oil biosynthesis genes

The evolutionary scenario of the oil biosynthesis system was evaluated with eight proteins of oleaster (Oeu000237.1; lipid transfer protein, Oeu042806.1; lipid transfer protein, Oeu014317.1; squalene biosynthesis, Oeu016163.1; oleosin biosynthesis, Oeu004694.1; Oleate desaturase, Oeu058547.1; Oleate desaturase, Oeu015599.1; Oleate desaturase and Oeu013924.1; Oleate desaturase) and their orthologous in 9 other plant species (*S. indicum*, *A. thaliana, E. grandis, G. max, O. sativa, P. persica,* *P. trichocarpa, S. tuberosum* and *V. vinifera*) retrieved from Phytozome (48) v9.1. and 15 other species (*Camelina sativa, Elaeis guineensis, Sorghum bicolor, Phoenix dactylifera, Cucumis melo, Ricinus communis, Malus domestica, Theobroma cacao, Medicago truncatula, Coffea canephora, Gossypium arboreum, Zea mays, Brassica rapa, Brassica napus* and *Arachis hypogaea*) obtained from NCBI GenBank in order to cover the commercial vegetable oil producing species as well as non oil-producing species (total number of species 25). Genes (proteins) used were from four different functional groups: i) oleate desaturase; ii) squalene biosynthesis; iii) oleosin biosynthesis; and iv) lipid transfer. Phylogenetic trees were constructed using the maximum-likelihood method with PhyML software (66). The best evolutionary model for the sequences was determined according to jModelTest (97) and considering Akaike's information criterion (AIC) (98). The selected model for both analyses was length of gaps (LG) with gamma (+G) distribution (99). Support for the nodes derived in these reconstructions was evaluated by bootstrapping, using 1,000 replicates (100). Eight oil-biosynthesis genes of oleaster and 24 other plant species were used for evolutionary analyses. As expected, the maximum-likelihood tree (derived from concatenated amino-acid sequences of selected oil-biosynthesis proteins) validated the evolutionary similarity between oleaster and sesame (Fig. S18).

S.4. Evaluation of oleaster genome

S.4.1. Oil biosynthesis genes and metabolic annotations unique to oleaster

Oil biosynthesis genes of *O. europaea* var. *sylvestris* were first clustered into 176 sub-classes, according to their functional annotations. Sub-groups having more than 10 members were listed as major clusters, in which cytochrome P450 monooxygenase (353) and AAA+ ATPase (199) were represented most. Expression patterns of important genes involved in oil biosynthesis, including fatty-acid synthases, elongases, desaturases, acetyl-CoA carboxylase, acyl dehydrogenase and β-ketoacyl-(acyl-carrier-protein) synthase (KAS) I to KAS III (total of 1,285 genes clustered into 37 groups) were comparatively analyzed. Four tissues (stem, leaf, pedicel and fruit) sampled at oil biosynthesis start (July) and end (November) seasons were analyzed. RPKM values of selected key genes related to oil biosynthesis were extracted from all RNA-seq datasets (four tissues in both July and November) (additional data file, see section S.6). A custom script was written to calculate log2 of RPKM averaged-value for all loci with the same gene functional-annotation or enzyme name. The resulting file was subsequently used as input for the heatmap construction by using an R <https://www.r-project.org> script available at <http://biotechvana.uv.es/nutritools/public/heatmaps_iterations>, applying three quartiles (0.3, 0.6 and 0.9) to differentially color expression patterns, reconstructing dendrograms for both rows and columns using the complete-linkage method with Euclidean-distance measure (101).

Oleaster genome functional annotation files were used to characterize oil biosynthesis-related genes, comparing them to eleven other plant genomes (*A. thaliana, E. grandis, G. max, O. sativa, P. persica, P. trichocarpa, S. tuberosum, V. vinifera, M. guttatus, U. gibba* and *S. indicum*), based on KEGG orthology (102) (additional data file, see section S.6, Dataset S.5-8). Lists of oil/lipid biosynthesis terms were used as searching criteria, in order to retrieve all matching oleaster gene IDs. Then, KEGG orthologs searches between oleaster and eleven other plant species were performed. That was related to oil biosynthesis genes extracted from KEGG orthology-annotation outputs of oleaster genome (additional data file, see section S.6, Dataset S.7). The list of genes related to the aforementioned processes was used. KEGG annotations of other species were downloaded from Phytozome. For simplicity's sake, only those with detectable orthologs in the oleaster genome were considered in each species. In other words, orthologs annotated via KEGG in the genomes of other species but not detected in oleaster were dismissed.

Interspecies KEGG data were processed using an *ad hoc* script to summarize information in two "per row" ways for "per KEGG ID" and "per pathway". Functional annotations and gene-expansion analyses for genes related to oil biosynthesis revealed that oleaster had the highest number of such annotated genes among all compared species. Moreover, most of them were expressed in at least one oleaster tissue. Comparing shared and unique KEGG annotations among oleaster and six other plant species (*A. thaliana, S. indicum, U. gibba, M. guttatus, G. max* and *P. trichocarpa*) showed 56 unique pathway hits in the former, such as K06130 lysophospholipase II.

S.4.2. Contribution of WGD to the expansion of *FabG, EAR, ACPTE* and *KASII* gene families

FabG (Beta-ketoacyl-ACP reductase), *EAR* (Enoyl ACP reductase), *KASII* (beta-ketoacyl-ACP synthase II) and *ACPTE* (acyl carrier protein (ACP)-hydrolase/thioesterase) are the major gene families that encode for enzymes playing a role in the oil biosynthesis pathway. In addition, they are important players to determine the oil composition, which changes from plant to plant. Preliminary studies showed that genetic manipulation on some of those members (silencing or over-expression) can change the plant oil composition (103, 104). During its evolution the oleaster genome has experienced two WGD events. After genome duplication, the vast majority of duplicated genes are eliminated due to the accumulation of deleterious mutations (105, 106), known as non-functionalization or pseudogenization. In opposite, the duplication event may extend the copy number of the gene, which either share the same function with its ancestral origin (subfunctionalization) or may gain a novel function (neofunctionalization) (106, 107). Either gene loss or retention after WGD can result in speciation and differentiation of a plant from its ancestral origin. Here, we revealed the impact of WGD to gene expansion of FabG, EAR and KASII families, which have 34, 52 and 7 members in the oleaster genome, respectively. Comparative phylogenetic analysis with different species (*A. thaliana* and *S. indicum*) was carried out and the contribution of WGD to gene expansion was evaluated between high oil producing plants (oleaster and sesame) and the other species (Table S17 and Fig. S29a-c). Acyl-ACP thioesterases (ACPTE) are divided into two groups, namely FatA and FatB, in which FatA is preferentially responsible for the hydrolysis of C18:1-ACP to C18:1 (oleic acid); and FatB catalyzes C16:0-ACP to C16:0 (palmitic acid) reactions (108). In the downstream process of the PUFA pathway, the FatA enzyme was observed as critical for the conversion of Oleoyl-ACP (C18:1-ACP) to oleic acid (C18:1), where it is represented by only two genes (*FatA-1* and *FatA-2*) in oleaster and one in sesame (Fig. S29d). Expression analysis revealed that *FatA-1* and *FatA-2* genes seem to be active in oleaster which concludes that two copies of *FatA* genes in oleaster and one in sesame control Oleoyl-ACP (C18:1-ACP) to oleic acid (C18:1) hydrolysis, thus it makes the *FatA* genes a target for biotechnological purposes.

S.4.3. *MADS-box* gene analyses

Minichromosome maintenance 1 (MCM1)-AGAMOUS (AG)-DEFICIENS (DEFA)-serum response-factor (SRF) (MADS)-box family proteins are transcription factors that play major roles in plant development, especially for floral organ and fruit development (see below). Identification, gene annotation, expansion and expression analyses were performed for MADS-box family proteins, potentially involved in both ripening and oil biosynthesis in the oleaster genome. A total of 109 Arabidopsis MADS-box proteins were used as a query for the identification of oleaster MADS box genes. In addition, candidate MADS-box members were checked for a transcription factor (TF)-specific domain, namely PF00319 SRF-type transcription factor (SRF-TF) (109). Some family members involve PF01486 K-box domain. Pfam database <http://pfam.xfam.org> v29.0 (110) was used to filter domains with E-value score $\leq 10^{-10}$. Oleaster MADS-box protein orthologs in eight other plant species (*A. thaliana, E. grandis, G. max, O. sativa, P. persica, P. trichocarpa, S. tuberosum* and *V. vinifera*) were retrieved from Phytozome. In addition, high-oil crop sesame MADS-box family protein were obtained from the *S. indicum* genome database (Sinbase) <http://www.ocri-genomics.org/Sinbase/index.html> (111, 112).

An *ad hoc* Python script was used to characterize sequences according to their length and sequence identities avoiding redundant matches and annotations. MAFFT (113) multiple-alignment tool was used to categorize sequences, by means of another *ad hoc* script. Evolutionary phylogeny analyses were conducted from 120 identified oleaster MADS-box sequences, using neighbor joining (NJ) method (114) in MEGA6 (86) and FigTree (version 1.4.2) <http://tree.bio.ed.ac.uk/software/Figtree>. Evolutionary distances were computed using the number-of-differences method (115) being the number of amino-acid differences per sequence. Parameters were set as: substitution, Poisson's model; data subset to use, complete deletion; and replication, for bootstrap analysis with 1,000 replicates. A heatmap analysis was conducted in order to differentiate expression patterns of MADS-box genes in eight tissues. RPKM value was mined from eight RNA-seq libraries and converted to log2. RStudio (v. 0.99.903) program with an R-script was used to draw the heatmap.

In total, 120 MADS-box genes/proteins were identified in the oleaster genome. These proteins segregated into two major groups as Type I and Type II. The first group included Mα, Mβ and Mγ clades, involving 22, 34 and 15 MADS-box proteins, respectively; The second group involved 49 proteins, corresponding to 41 "MADS-intervening keratin-like and C-terminal" (MIKC)-type sequences and eight Mδ sub-group sequences (Table S19 and Fig. S30). All oleaster MADS-box proteins included at least one SRF-TF (PF00319) domain. In addition, K-box domain (PF01486) was found in 11 MADS sequences, belonging to MIKC group (Fig. S30a). Cluster analyses showed that at least 37 oleaster MADS-box proteins shared consensus sequences with selected species. Notably, 11 clusters belonged to high-oil producing plants (oleaster and sesame). Black branches and clusters in dendrogram were specific of oleaster. Cluster sequences are shown in color gradation from dark red (nine species) to deep pink (two species) (Fig. S30b). On the other hand, most of oleaster MADS-boxes among selected species were species-specific or had unique function. These results can help to understand functions of MADS members in oleaster, especially in fruit development/ripening and oil biosynthesis mechanisms.

In recent years, many MADS-box genes were functionally verified as being involved in fruit ripening of several plants like tomato (116, 117), strawberry (118), apple (119), banana (120) etc. In general, over-expression of ripening-associated MADS-box family members cause delayed fruit ripening, except *Solanum lycopersicum* MADS-box 1 (SlMADS1) (120, 121). Genome analyses and identification of MADS-box genes in oleaster exhibited valuable data, revealing critical information in relation to oil biosynthesis and fruit-ripening bioprocesses. Thus, some MADS-box transcripts were highly expressed in July to November fruits (Fig. S30a). For instance, November-fruit Oeu012464.1 transcribed six times more (log2) than July one. Also, Oeu002302.1 and Oeu046544.1 were abundantly expressed (at least three times, log2) in November fruit. Interestingly, oil production takes place during the fruit-maturation stage, especially in the ripening period, which is November for oil-bearing plants like oleaster. Highlighted MADS-box TFs can be potential genes/proteins involved in oleaster oil biosynthesis and fruit ripening progression. Also, protein annotation and cluster analyses revealed that Oeu012464.1 was grouped in Cluster 2 with five other species and Oeu002302.1 was in Cluster 12 with three other species. Oeu046544.1 has only one ortholog member (SIN_1025213 or SiMADS52), which belongs to high oil producing crop sesame (Fig. S30b). Thus, these three MADS-box genes are potential regulators of oil production and fruit-ripening process in oleaster (further experiments are required to evaluate such a hypothesis).

S.4.4. Ripening genes of oleaster

S.4.4.1. Genes involved in ripening

Lists of gene names related to ripening were retrieved from scientific literature and gene/protein databases available at NCBI. Oleaster genes involved in ripening were listed (additional data file, see section S.6, Dataset S.9). A total number of 10,976 genes were found to be involved in such a process. Some of them were highly represented, including gene families such as protein serine/threonine kinases, zinc finger transcription factors (CCHC-type), oxidoreductases and hydrolases.

S.4.4.2. Ripening genes and metabolic annotations unique to oleaster

Fruit ripening is a complex phenomenon, summing various metabolic, structural and physiological processes, including hormonal regulation, metabolite synthesis, accumulation of volatiles, sugar and other carbohydrate metabolism, and nutrient content changes (122). Different developmental stages of oleaster ripening may modulate the content and yield of oil, antioxidants like phenolics, and fruit (table olive) organoleptic properties. KEGG ortholog search, metabolic annotation, histogram, heatmap and Venn diagram analyses were performed as indicated in section S.4.1.

Ripening-related gene annotations were performed, resulting in 10,976 gene annotations in oleaster (additional data file, see section S.6, Dataset S.9). Transcription factors such as MADS-box, zinc finger and basic leucine-zipper (bZIP) were identified as involved gene families. This supports the importance of transcriptional regulation. To further analyze ripening-related pathway summarization, KEGG ortholog search (102) and metabolic annotation for ripening-related genes were performed in 10 plant species including oleaster. The latter showed the highest gene numbers for corresponding metabolic annotations, such as folding, sorting and degradation (1,763), carbohydrate metabolism (917) and signal transduction (717) (Fig. S19). Interestingly, genome-wide comparisons of metabolic annotations corresponding to ripening-related genes with five other plant species (*A. thaliana, S. indicum, E. grandis, O. sativa* and *P. trichocarpa*) revealed that 283 of 1,283 KEGG annotations were unique to oleaster (Fig. S19). Additionally, comparison in expression pattern of selected ripening-genes showed some genes with similar expression patterns, like zeaxanthin epoxidase and E3 protein ligase RFWD2 (Fig. S20a). Interestingly, fruit samples exhibited similar expression patterns as pedicel tissues at different development and maturation stages (Fig. S20b). It is well defined that MADS-box family proteins are involved in transcriptional regulation of ripening processes (122). Indeed, a tight link was found between oil biosynthesis and ripening when MADS-box family proteins were taken into consideration for such processes. This involves common genes and metabolic pathways, such as lipid metabolism and secondary-metabolite biosynthesis.

S.4.5. Carbohydrate and lipid-metabolism genes

Carbohydrates represent important forms of stored energy. As such, they are metabolized for a number of biological processes, including oil biosynthesis in plants. Long-chain polysaccharides are formed from glucose by a number of metabolic processes, including breakdown and interconversion. Ortholog comparison for carbohydrate and lipid metabolism level with eight other plant species (*A. thaliana, E. grandis, G. max, O. sativa, P. persica, P. trichocarpa, S. tuberosum* and *V. vinifera*) were performed using KEGG orthology-annotation outputs of the oleaster genome. Analyses were carried out as described in section S.4.1 showing KEGG ID and pathway.

A total of 2,996 genes involved in carbohydrate metabolism encoded by oleaster genome were metabolically annotated. Compared to other plant species, such genome had the highest number of genes related with pathways of carbohydrate metabolism, such as inositol phosphate (138) and galactose (93) (Fig. S21a). On the other hand, interestingly, 34 of 207 KEGG annotations related with carbohydrate metabolism were unique for oleaster, in comparison with four other plant species (*A. thaliana, E. grandis, O. sativa* and *P. trichocarpa*) (Fig. S21b). A number of lipid types are also produced, modified and stored by plants for many purposes, such as deposition and storage of energy sources, signaling molecules and hormonal homeostasis, and construction of cellular structures. Oleaster also synthesizes several lipids other than fatty-acids, such as arachidonic acid and sphingolipids. Indeed, pathway annotations based on KEGG orthology identified 1,266 genes for lipid metabolism in such species. Being an oil-crop, oleaster had higher numbers of annotated genes for lipid metabolism, compared to other plant species like *G. max* (822) (Fig. S22a). Interestingly, out of 147 KEGG annotations related with lipid metabolism, 45 unique oleaster pathways were found (Fig. S22b).

S.4.6. Secondary metabolite biosynthesis gene analyses

Biosynthesis of active compounds in olive fruit and leaves includes antioxidant phenolics (oleuropein, hydroxytyrosol, alpha-tocopherol or vitamin E, carotenes, etc). They have key relevance in medicine and cosmetics, being main micronutrients of the healthy Mediterranean diet (123). Oleuropein, a secoiridoid-type metabolite with pharmacological effects (including anticancer, antiinflammatory and hypoglycemic activities), is an Oleaceae-specific compound, found in the fruit mesocarp and leaves as a phenylethanoid-pathway product (124). Therefore, the oleaster genome was mined for brassinosteroid- (BR) signaling, shikimate and carotenoid pathways, as major secondary metabolite biosynthesis processes, as well as phenylethanoids (PE) biosynthesis metabolism producing oleuropein (Fig. S23a-d). Interestingly, oleaster had the highest observed number of coumarate 3-hydroxylase C3H (n = 89), involved in PE pathway (total n = 259 genes), compared to other sequenced plants.

Different secondary metabolite pathway genes, including terpene synthesis (TPS), brassinosteroid (BR) signaling, shikimate, carotenoid, phenylethanoids (PE) and N-methyltransferases (NMT), were identified in oleaster genome. Genomic datasets from 10 plant species (*A. thaliana*, *E. grandis*, *G. max*, *O. sativa*, *P. trichocarpa*, *P. persica*, *S. indicum*, *S. tuberosum*, *T. cacao* and *V. vinifera*) were used for secondary-metabolite pathway-related protein mining. Two different approaches were applied to identify gene families of selected pathway proteins in 10 different plant genomes. Firstly, BLASTP searches were performed against the Phytozome database using default parameters. Secondly, Hidden Markov Model (HMM) profiles of related domains were selected from the Pfam database. Obtained query sequences were then searched against the oleaster proteome using BLASTP. Hits with E-value score above

1×10^{-10} were retained as candidate genes responsible for secondary-metabolite biosynthesis. The presence of Pfam-identifier numbers was checked again, according to conserved protein-domains. Comprehensive phylogenetic analyses were performed to reveal relationships between secondary-metabolite gene families of oleaster and other plant species. All amino-acid sequences were imported into MEGA7 (125). Multiple-sequence alignments were conducted using MUSCLE. Maximum-likelihood trees were built with Jones-Taylor-Thornton (JTT) model (126) and 1,000 bootstrap replicates. Then, dendrogram was displayed using interactive Tree of Life (iTOL) version 3 <http://itol.embl.de> (127).

RNA-seq libraries from different oleaster tissues and time points were used for determination of gene-expression levels of all secondary metabolites gene families. Firstly, RPKM values of oleaster genes were calculated from these libraries and converted to log2. Then, two-way hierarchical clustering heatmaps were generated. Finally, all data were transferred into PermutMatrix software <http://www.atgc-montpellier.fr/permutmatrix> (128) visualized and further analyzed with heatmaps. Secondary-metabolite-synthesis gene and transcript analyses were performed. TPS, NMT, BR and PE pathway-related genes, carotenoid-biosynthesis and shikimate-pathway genes were annotated. A total of 51 and 15 genes were identified for TPS and NMT, respectively. Brassinosteroids play a crucial role regulating physiological and developmental processes during the whole life of plants. They are plant steroid-hormones which interact with other signaling networks to regulate diverse physiological-processes and stress responses (129). A high number of brassinosteroid genes, including brassinosteroid hydroxylase/oxidases [cytochrome P450 (*CYP*); 316 genes], brassinosteroid 23-O-glycosidase (*UGT73C5*; 154 genes) and steroid 5-alpha-reductase (*DET2*; four genes) were annotated in the oleaster genome. In addition, a total of 316 oleaster *CYP* genes were also phylogenetically clustered into 16 different groups. Their expression levels also showed different expression patterns, based on different tissue and sampling time (Fig. S23a).

Shikimate or shikimic-acid pathway is responsible for biosynthesis of aromatic amino acids (AAA; phenylalanine, tyrosine and tryptophan) in bacteria, fungi, algae and plants. It also provides different precursors for production of natural products, such as pigments, alkaloids, hormones and cell-wall components (130). Such AAA are essential components of animal diets, including humans. A total of 69 shikimate-pathway genes were discovered in oleaster genome. Prephenate aminotransferase (*PAT*; 23 genes) and arogenate dehydratase (*ADT*; 26 genes) had the highest copy numbers in the oleaster genome. In contrast to them, chorismate mutase (*CM*) and 3-dehydroquinate synthase (*DHS*) genes were single copy. Transcriptome analyses showed that *CM* and *DHS* had tissue-specific expression patterns. For example, *DHS* gene-expression level increased in fruit and leaf samples of July and November time points, whereas expression of the *CM* gene rose in petiole and stem samples of both time points (Fig. S23b).

Carotenoids are a group of isoprenoid molecules, generally regarded as pigments, which participate in light harvesting and photoprotection against excess light in plants. They are also responsible for biosynthesis of precursors for the production of apocarotenoid hormones (abscisic acid and strigolactones) (131). Among 2,243 carotenoid-pathway genes from 10 plant species, a total of 193 oleaster specific genes, which catalyze carotenoid metabolism, were identified in the oleaster genome. Through this pathway, zeta-carotene desaturase (*ZDS*; 10 genes), lycopene epsilon-cyclase (*LUT2*; three genes), lycopene beta-cyclase (*LYC*; two genes), carotene epsilon-monooxygenase (*LUT1*; 61 genes) and beta-hydroxylase (*BH*; 23 genes) catalyze reactions, which result in the production of lycopene, α- and β-carotene, lutein and zeaxanthin, respectively. Their gene expression exhibited various patterns in different tissues and sampling times (Fig. S23c).

PE is a complex pathway composed of a combination of different pathways. It starts with shikimate acid, continues with the production of tyrosine and phenylalanine, and finally ends with accumulation of cyclohexylethanol derivatives and phenylethanoid glycosides. They are then converted into salidroside and oleuropein through complex and unidentified steps. In total, 259 genes active in the phenylethanoid pathway were identified in the oleaster genome. Among them, the largest gene families were uridine-diphosphate glucose (*UDP*)-glycosyltransferases (*UGT*; 37 genes) and coumarate 3-hydroxylase (89 genes) (Fig. S23d). Gene expression measurements of all secondary-metabolite pathways are shown in Fig. S23a-d.

S.4.7. Alternate bearing related genes analyses

Alternate bearing (AB; periodicity) is a curious phenomenon by which some trees produce abundant flowers and fruits one year, yet low or none in the next harvesting season, known as on/off years, respectively. It is a hormonal process triggered by the amount of carbohydrate reserves, as we have described in the wild olive tree (132, 37, 133, 134, 132). Genes involved in the AB mechanism were identified from oleaster and eleven other plant species, by following the procedure indicated in section S.4.5. A total of 275 genes related to the AB process were annotated and analyzed in oleaster (Table S18).

S.4.8. Identification of transcription factors in the oleaster genome

In order to determine oleaster transcription-factors, we started by retrieving TF protein-sequences from the proteomes of *A. thaliana*, *S. indicum*, *S. tuberosum* and *V. vinifera*. Plant Transcription-Factor database (PlntTFDB) v3.0 <http://planttfdb.cbi.pku.edu.cn> (135) and Phytozome were used to obtain query sequences for each species. They were aligned to the oleaster protein sequences with BLASTP. Only hits with an E-value score $<10^{-10}$ were selected as TF candidates. Characteristic conserved-domains for each TF family were searched according their Pfam identifier numbers, which were determined according to transcription-associated proteins (TAP) rules. E-value score $<10^{-10}$ was set as the selection criteria for complete domains. The presence of complete TF-specific domains and characteristic residues was also checked by using the Clustal Omega multiple-sequence alignment program for proteins <http://www.ebi.ac.uk/Tools/msa/clustalo>. Oleaster TFs were compared to ten sequenced plant species, which comprised model dicots (*A. thaliana* and *V. vinifera*), model monocot (*O. sativa*) and oil-bearing plants (*S. indicum*, *G. max* and *T. cacao*). In total, 44 TF families were scanned and listed in Table S19.

In general, a large number of oleaster TFs were found to be expanded in comparison to the model oil-bearing plants *S. indicum* and *T. cacao*. Interestingly, they were relatively low in other species like soybean, rice and poplar. The number of TF family-members was quite constant between databases used and other publications. However, some genomes exhibited variable number of TF members, in which case the latest updated studies were used. TFs are known to be regulating gene expression in eukaryotes. They are involved in different pathways, including lipid biosynthesis. For instance, basic helix-loop-helix (bHLH) TF in sesame (136); basic leucine zipper (bZIP) (137) and DNA-binding with one zinc finger (DOF) TF in soybean (138) play key roles in regulation of lipid biosynthesis. Thus, overexpression of microalgae bHLH TF increases lipid production by enhancing growth and nutrient uptake (139). Also, overproduction of a DOF-type TF increased lipid content of *Chlamydomonas reinhardtii* and *Chlorella ellipsoidea* (140, 141). In this context, oleaster TF genes can be candidate targets for engineering valuable metabolites and lipids.

S.4.9. Genome-wide protein-family analyses

Oleaster proteins together with proteins of eight other plant species (*A. thaliana, E. grandis, G. max, O. sativa, P. persica, P. trichocarpa, S. tuberosum* and *V. vinifera*) were retrieved from Phytozome. In short, all sequences were merged into a single file and processed for cluster analyses using Prototype-Simulate-Interact Cluster Database at High Identity with Tolerance PSI-CD-HIT (142) with a similarity threshold of 30%. Clustering results were integrated into a single CSV file, together with ID and definition information from InterPro (143) system of classification using GPRO. They were summarized in two different per-row ways: per Cluster ID and per InterPro definition. Zinc finger, the really interesting new gene (RING)-type (344), myeloblastosis (Myb) domain, DNA-binding (333) and APETALA2/ethylene-responsive element (ERE) binding factor (AP2/ERF) domain (283) were the most represented protein families in the oleaster genome. Comparing to eight other plant genomes, "nucleotide-binding adapter shared by apoptotic protease-activating factor 1 (APAF-1), disease resistance (R) gene products and cell-death activator (CED-4)" (NB-ARC) domain (IPR002182)-encoding genes were highly represented (290) in the oleaster genome. Moreover, terpene-synthase domain (IPR005630 and IPR001906)-encoding genes involved in secondary metabolite biosynthesis (40 genes), peptidase C48 [small ubiquitin-like modifier (SUMO)/sentrin/ubiquitin-like protein 1 (UBL1)] (IPR003653) involved in ripening (151 genes) and conserved-sequence motif "aspartic acid-histidine-histidine-cysteine" (DHHC)-type palmitoyltransferase zinc-finger-domain-encoding genes (153) involved in oil biosynthesis pathways were also found in higher number in the oleaster genome compared to other plant species (Fig. S24).

S.4.10. Disease resistance gene analyses

Main disease-resistance family members of “nucleotide binding-site leucine-rich repeat” (NBS-LRR) oleaster genes were also annotated with their expression and expansion analyses, by following the method explained in section S.4.5. Pfam disease-resistance protein-domains of nucleotide-binding site (NBS; PF00931), N-terminal toll/interleukin-1 receptor (TIR; PF01582) and leucine-rich repeat (LRR; PF00560, PF07723, PF07725, PF12799, PF13306, PF13516, PF13855 and PF14580) were used. HMM-Pfam search was carried out by using a hidden Markov-model-based sequence (mer) alignment tool HMMER v3 <http://hmmer.org> (144). Candidate oleaster NBS-LRR proteins were sorted into six groups: TIR-NBS and TIR-NBS-LRR (TIR type); NBS, coiled-coil (CC)-NBS, NBS-LRR and CC-NBS-LRR (non-TIR type). N-terminal CC motifs cannot be identified by using Pfam analyses. A hidden Markov-model-based program to identify putative coiled-coil domains in protein sequences (MARCOIL) (145) with threshold score of 90 was used. Additionally, results were validated with Paircoil2 (146) program, with a P-score cut-off of 0.03 (147). Such application predicts parallel coiled-coil-fold in peptide sequences using pairwise-residue probabilities.

A total of 290 NBS-LRR R-genes were identified in oleaster genome. NBS-LRR proteins generally separated into two distinct groups; namely, TIR and non-TIR types. Most woody and herbaceous plants contain both of them. Interestingly, the oleaster genome does not encode any TIR-type member, as also found in the other oil-crop genome (*S. indicum*) (92). Although genome-wide scale (PCR-based) studies have reported the absence of R-genes in sugar beet (148) and some monocots (149), the lack of TIR-type R-genes in other plant species has not been previously published, as far as we know. It would be interesting to further ascertain the absence or reduction of such genes in genome-wide analyses of other oil crops (Table S20 and Fig. S25).

S.4.11. sRNA mapping on *FAD2* genes

To map small RNA sequences we used the sRNA libraries generated by our group (37) including from ripe fruit collected from the November season tissue. Before mapping we took the 10 kb regions of corresponding *FAD2* genes as follows *FAD2-1* gene, Oeu013924.1, location; chr4 starting from 15,922,000, *FAD2-2* gene, Oeu058547.1, location; chr22 starting from 3,265,002, *FAD2-3* gene, Oeu033739.1, location; chr17 starting from 707,001, *FAD2-4* gene, Oeu007766.1, location chr9 starting from 12,560,000 and *FAD2-5* gene, Oeu061755.1, location chr3 starting from 23,360,987 (Fig. S26). Here, we showed that the up and down-stream regions of *FAD2-1, FAD2-2, FAD2-4* and *FAD2-5* genes are rich in TEs but the TE density is relatively low in the 10 kbp region of the *FAD2-3* gene. Then, the sRNA reads were mapped to those 10 kb regions via CLC Genomics Workbench (v7) with default parameters. The sRNAs mostly mapped on the 5'-UTR regions of the *FAD2* genes. The possible siRNA binding sites were marked on the UTR regions (Fig. S27a-e). On the other hand, nearly no sRNA mapping peaks were detected on the 5'-UTR region of the FAD2-3 gene (Fig. S27c). Then, we searched the siRNA consensus sequence using the siRNA peaks on the FAD2 5'-UTR regions. We concluded that a siRNA consensus sequence (5'-CTT NAA TCA ANN ACA ACC CNA) binds to the FAD2-1, FAD2-2, FAD2-4 and FAD2-5 transcripts but cannot bind to FAD2-3 due to the presence of 12 additional nucleotides at the binding site (Fig. S27f).

S.5. Genome browser development

The "Olive Genome Browser" was built <http://h3abionet.fso.ump.ma/cgi-bin/gb2/gbrowse/olea_europea> using the MySQL 5.1.73 relational database <http://www.mysql.com> and the Perl scripting language on CentOS <https://www.centos.org>. The Apache web server <https://www.apache.org> was used to handle requests from user's web browsers. It validates users and sends requests to a combination of database and interactive web pages for manipulating and displaying annotations on genomes (GBrowse) <http://gmod.org/wiki/GBrowse>. Perl modules in the latter analyze and validate queries, sending them to the MySQL database, which responds to them. Requested data are then transferred to the Apache web server and translated into graphical representations, where users can visually explore results via the intuitive interface of web browsers. The required user id is: "badad" and password is: "&Rabat%2b!" to reach the database.

S.6. Dataset information

The genome assembly was uploaded to NCBI WGS (accession number: MSRW00000000).. The RNA-seq transcriptome datasets were uploaded into NCBI Sequence Read Archive with submission number SUB2036285 (SRR4473639, SRR4473641, SRR44742, SRR4473643, SRR4473644, SRR4473645, SRR4473646 and SRR4473647).

Additionally, the full dataset and all additional data files and tables can be downloaded from <http://olivegenome.org/downloads>. The required password is: "ibg" to reach the datasets. The Phytozome link can also be used to reach the datasets <http://portal.nersc.gov/dna/plant/annotation/wild_olive/>. As well as from ORCAE < http://bioinformatics.psb.ugent.be/orcae/>

Downloadable Files

GeneOnthology_Olive.xlsx; Gene ontology results of oleaster genome, including, annotation descriptions, GO terms, KEGG pathways and map IDs (Dataset S.5).

IPRscan_Olive.xlsx; InterProScan outputs of oleaster genome annotations (Dataset S.6).

KEGG_Orthology_Olive; KEGG orthology annotation outputs of oleaster genome (Dataset S.7).

GeneOnthology_Sindicum.xlsx; Gene onthology results of *S. indicum* genome (Dataset S.8).

Oil_Biosynthesis_Genes_Olive.txt; Genes involved in oil biosynthesis of oleaster (Dataset S.2).

Oil_Biosynthesis_Genes_Sindicum.txt; Genes involved in oil biosynthesis of *S. indicum* (Dataset S.3).

Ripening_genes_olive.txt; Genes involved in ripening process of oleaster (Dataset S.9).

Figures

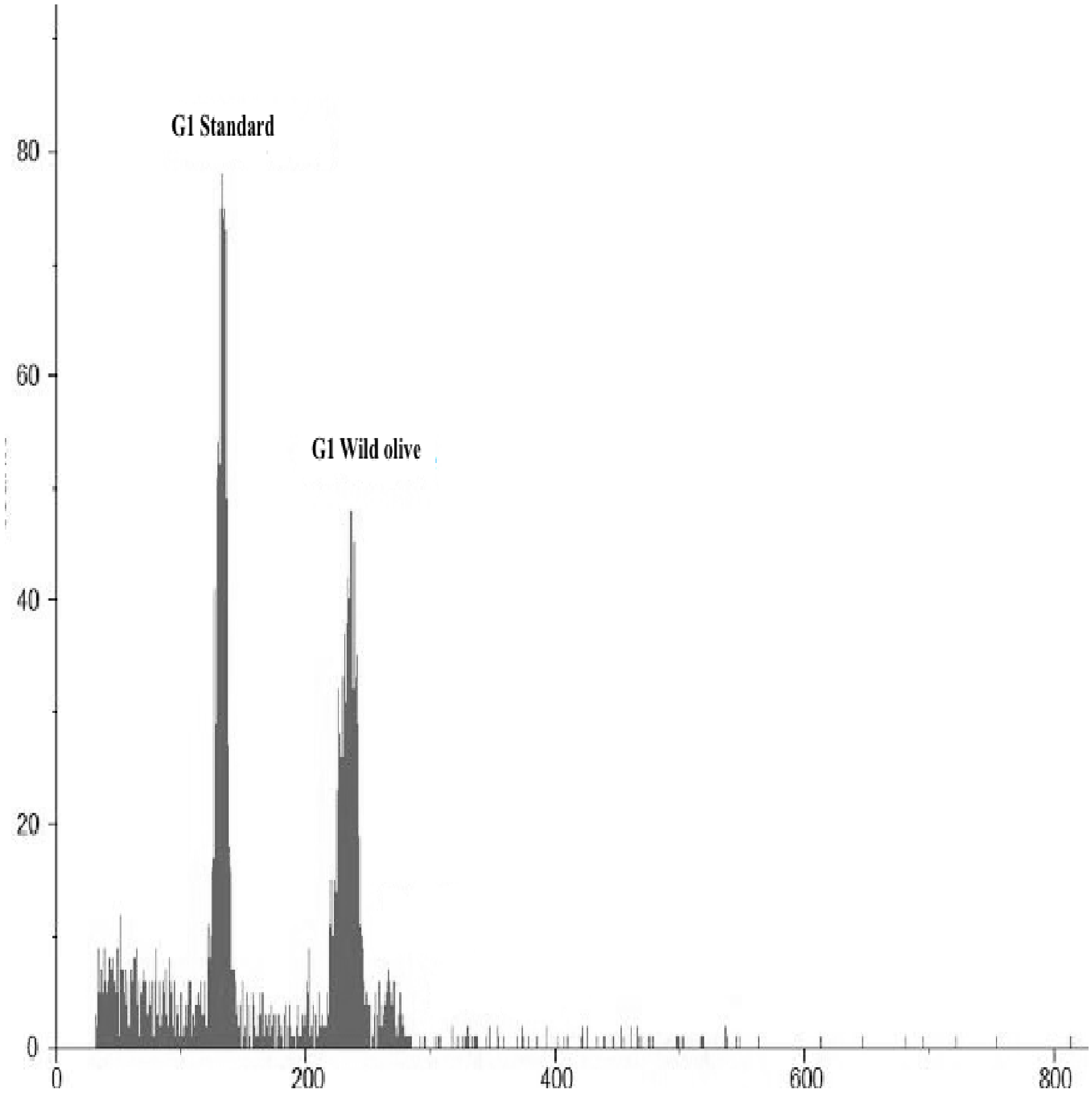


Figure S1. Oleaster genome size estimation. Histogram of relative nuclear-DNA content using CyStain PI Absolute P. Chickpea (*Cicer arietinum*) was used as internal standard. The G1 peak of chickpea was detected on channel 132 (740 Mbp for 2C); hence, the 2C DNA amount of oleaster is 2.91 Gb. X-axis represents the relative content and Y-axis corresponds to the number of events in the graph.

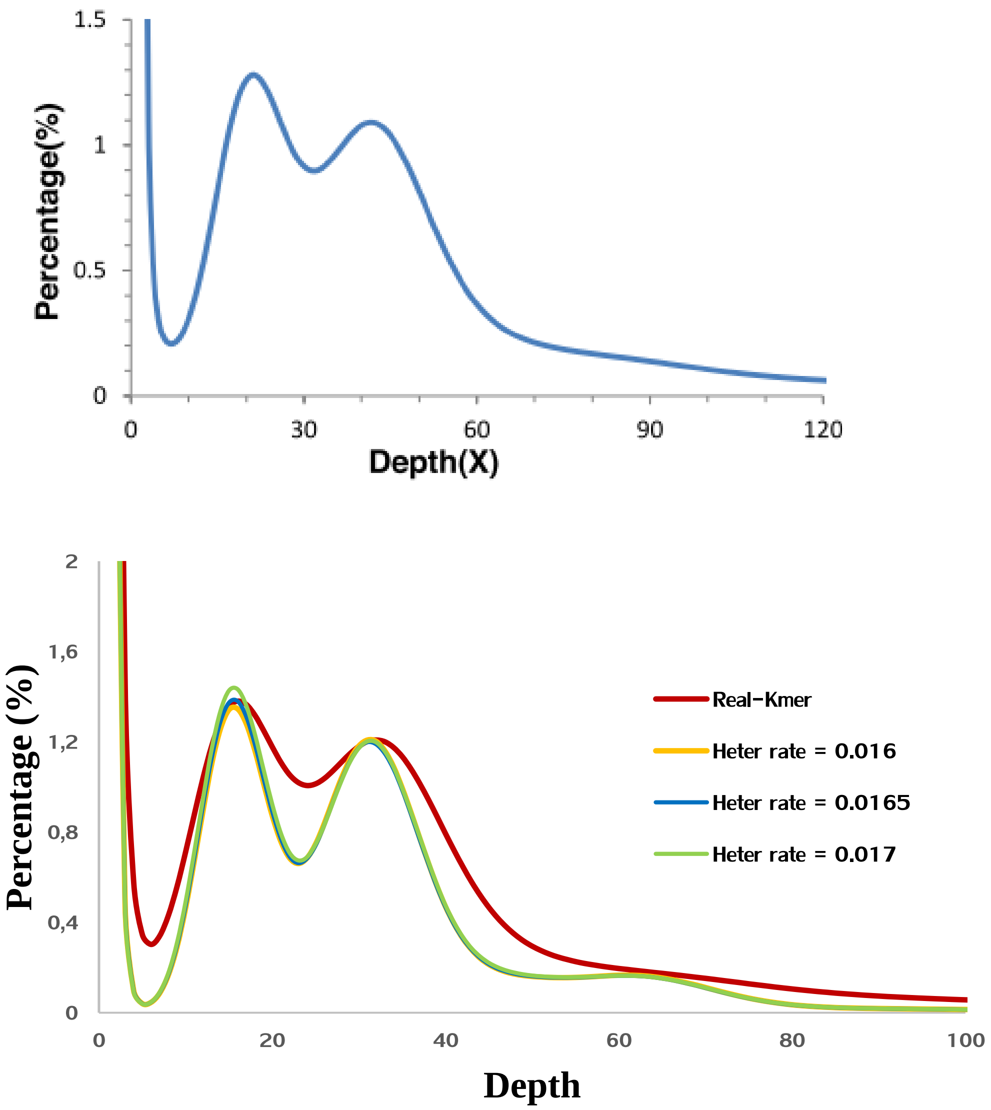


Figure S2. **Genome size estimation (17-mer). A)** The 17-mer frequency distribution derived from the sequencing reads was plotted. A total of 68.92 Gbp were used for *k*-mer analysis. Abscissa corresponds to depth (X) and ordinate is the percentage, which represents the frequency at that depth divided by the total frequency of all depths, multiplied by 100. The 17-mer distribution should obey the Poisson theoretical distribution, without considering sequence error, heterozygosis and repeat rates of the genome. In practice, the low depth of *k*-mer frequency should take up a large proportion in the actual data, due to sequence error. Likewise, heterozygosis may generate a subpeak associated to the main peak. Additionally, repeats can cause repeat peaks at multiple integers of the main peak. **B)** The heterozygous rate was estimated by simulating the *k*-mer distribution of the heterozygous sequence.

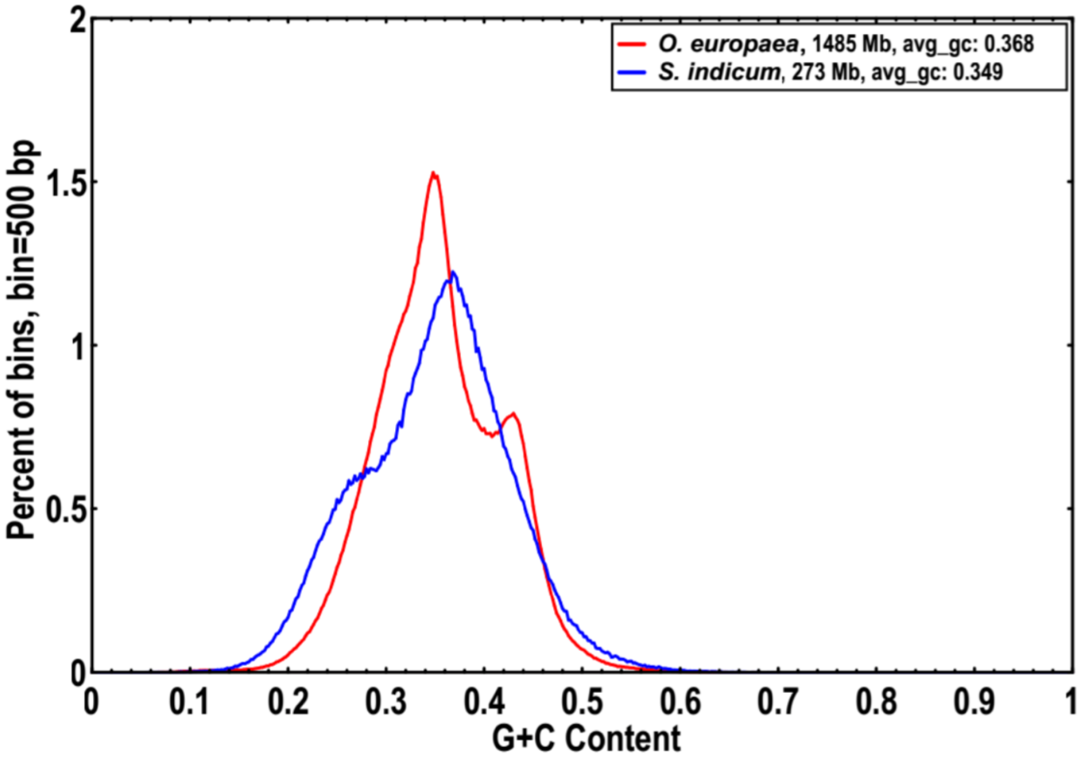


Figure S3. **GC-content distribution of *O. europaea* var. *sylvestris* and *S. indicum* genomes.** Abscissa represents GC content and ordinate corresponds to the proportion of the bins number divided by the total windows, multiplied by 100. Graph was obtained using a 500 bp sliding window (with 250 bp overlaps) along the genome.

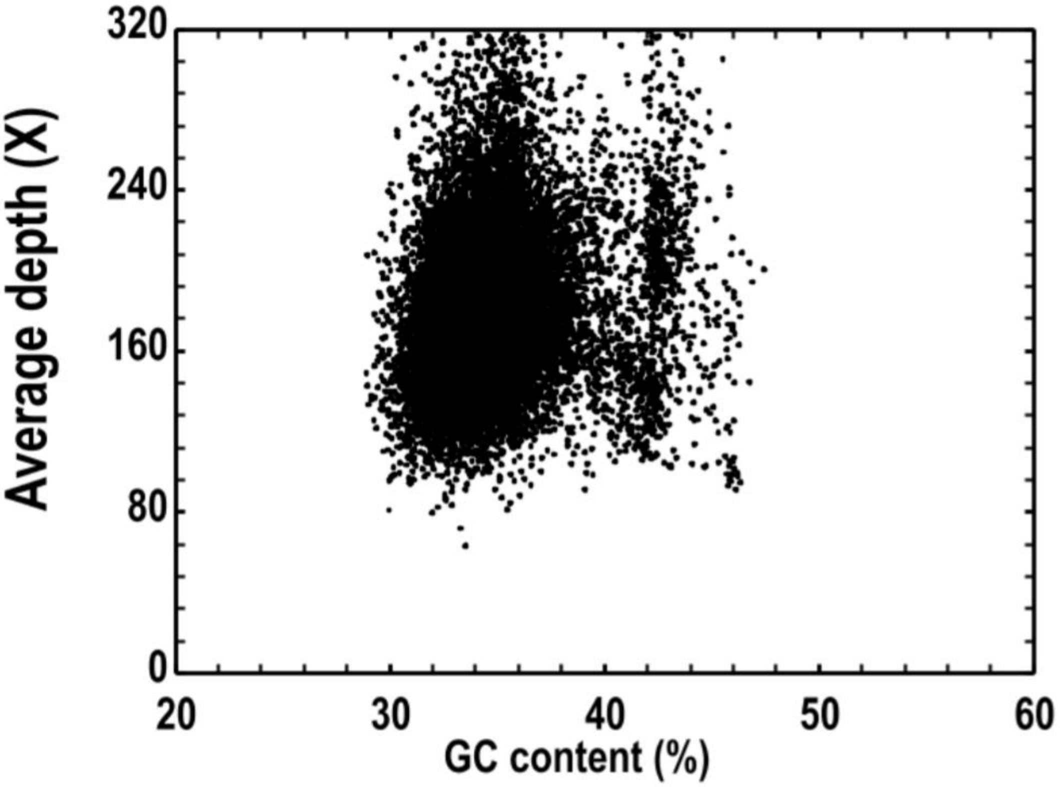


Figure S4. **GC content and sequencing-depth analyses.** Abscissa represents GC content and ordinate shows average depth. Plot was obtained with 50 kbp non-overlapping sliding windows, calculating GC content and average depth among windows.

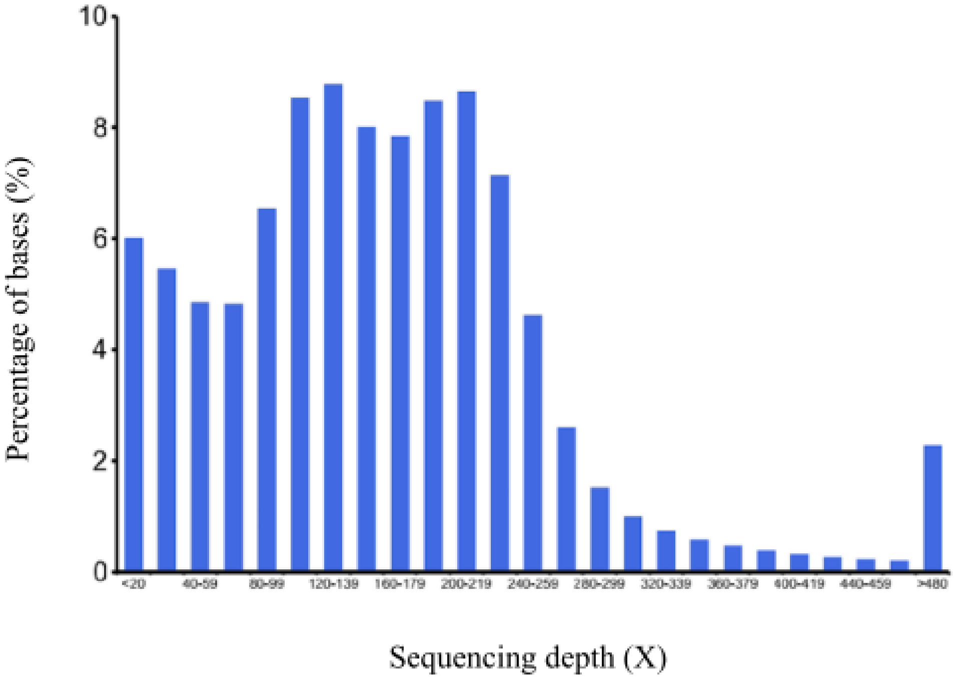


Figure S5. Sequence-depth distributions. Abscissa shows depth and ordinate corresponds to proportion of base number divided by total bases, multiplied by 100. Filtered reads were aligned onto the assembly genome sequence using SOAP. Then the percentage of bases with different depth in the genome was calculated.

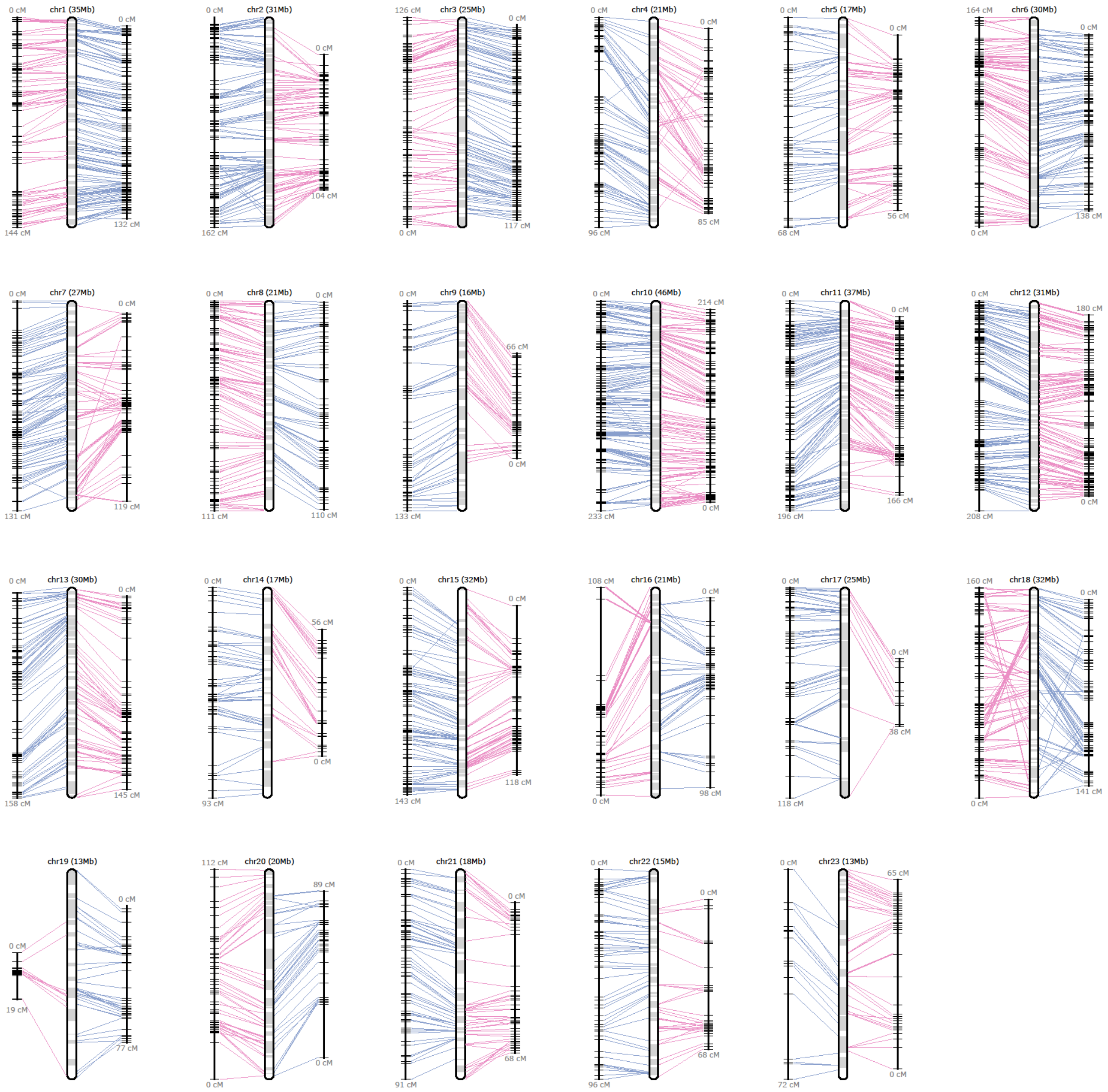


Figure S6. Linkage groups of the oleaster genome. Physical map (middle bar) was anchored to genetic maps. Left and right bars correspond to molecular-marker positions in genetic maps, generated by Genotyping-by-Sequencing (GBS).

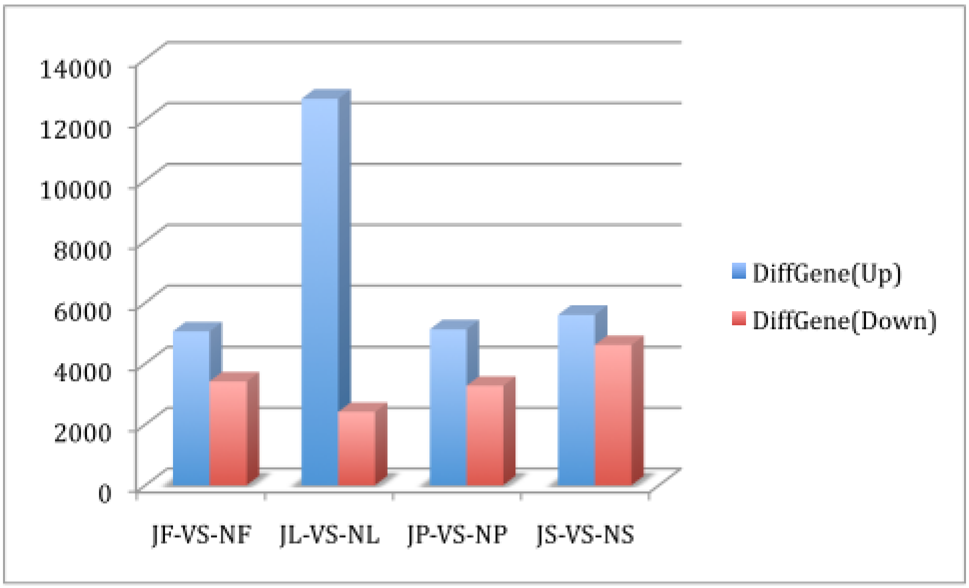


Figure S7. Differential expression levels of genes in RNA-seq libraries. Expression levels of genes were compared between libraries, based on RPKM values. Each library had at least 81% of genome map with 41,559 genes (see Table S9). J: July, N: November, F: Fruit, L: Leaf, P: Pedicel and S: Stem.

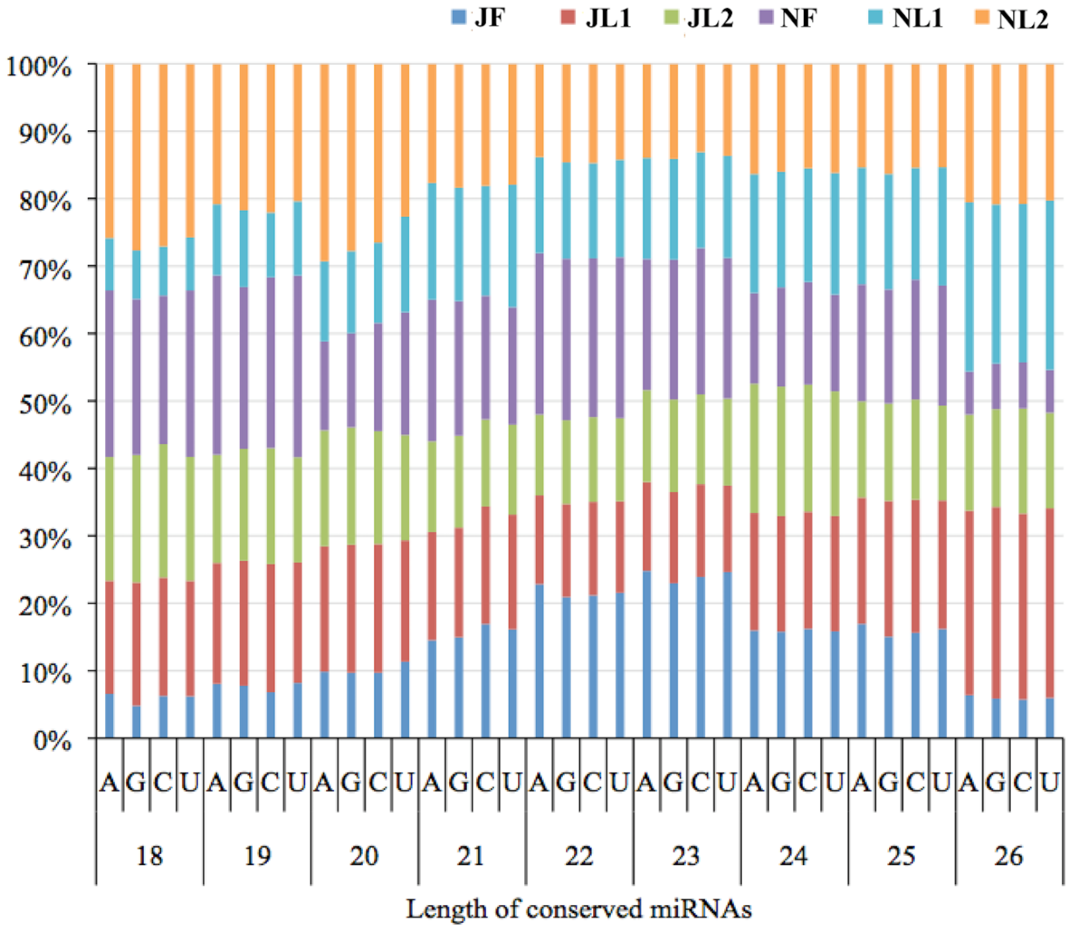


Figure S8. Nucleotide composition of oleaster. Drawing shows conserved miRNA of 18 to 26 bases. Nucleotide distributions of 18 to 26 b reads were similar in the six libraries. J: July, N: November, F: Fruit, L: Leaf.

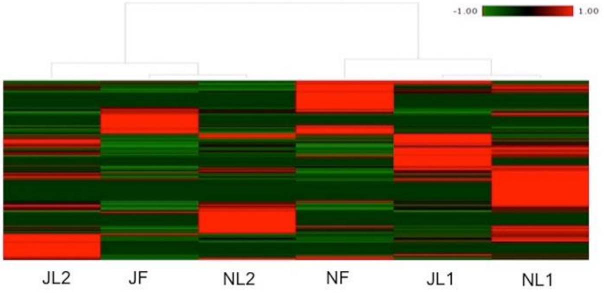
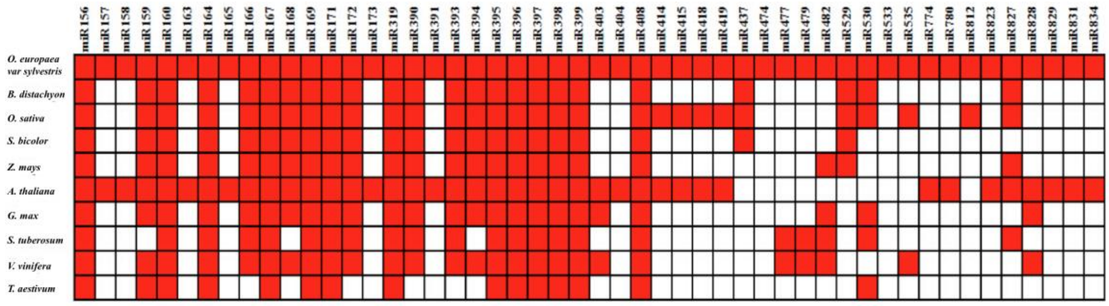


Figure S9. Heatmap of conserved miRNA expression in six oleaster samples. A total of 334 (67.8%) conserved miRNA families were differentially expressed in the six oleaster libraries (*p*-value ≤0.01). J: July, N: November, F: Fruit, L: Leaf.



	miR156	miR157	miR158	miR159	miR160	miR163	miR164	miR165	miR166	miR167	miR168	miR169	miR171	miR172	miR173	miR319	miR390	miR391	miR393	miR394	miR395	miR396	miR397	miR398	miR399
O. europaea var sylvestris	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
B. distachyon	🟥			🟥	🟥		🟥		🟥	🟥	🟥	🟥	🟥	🟥		🟥	🟥		🟥	🟥	🟥	🟥	🟥	🟥	🟥
O. sativa	🟥			🟥	🟥		🟥		🟥	🟥	🟥	🟥	🟥	🟥		🟥	🟥		🟥	🟥	🟥	🟥	🟥	🟥	🟥
S. bicolor	🟥			🟥	🟥		🟥		🟥	🟥	🟥	🟥	🟥	🟥		🟥	🟥		🟥	🟥	🟥	🟥	🟥	🟥	🟥
Z. mays	🟥			🟥	🟥		🟥		🟥	🟥	🟥	🟥	🟥	🟥		🟥	🟥		🟥	🟥	🟥	🟥	🟥	🟥	🟥
A. thaliana	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
G. max	🟥			🟥	🟥		🟥		🟥	🟥	🟥	🟥	🟥	🟥		🟥	🟥		🟥	🟥	🟥	🟥	🟥	🟥	🟥
S. tuberosum	🟥				🟥		🟥		🟥	🟥		🟥	🟥	🟥		🟥	🟥		🟥		🟥	🟥	🟥	🟥	🟥
V. vinifera	🟥			🟥	🟥		🟥		🟥	🟥	🟥	🟥	🟥	🟥		🟥	🟥		🟥	🟥	🟥	🟥	🟥	🟥	🟥
T. aestivum	🟥			🟥	🟥		🟥			🟥		🟥	🟥			🟥					🟥	🟥	🟥	🟥	🟥

	miR403	miR404	miR408	miR414	miR415	miR418	miR419	miR437	miR474	miR477	miR479	miR482	miR529	miR530	miR533	miR535	miR774	miR780	miR812	miR823	miR827	miR828	miR829	miR831	miR834
O. europaea var sylvestris	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
B. distachyon			🟥					🟥					🟥	🟥							🟥				
O. sativa			🟥	🟥	🟥	🟥	🟥	🟥					🟥	🟥		🟥			🟥		🟥				
S. bicolor			🟥					🟥					🟥												
Z. mays			🟥									🟥	🟥								🟥				
A. thaliana	🟥	🟥	🟥	🟥	🟥	🟥	🟥										🟥	🟥		🟥	🟥	🟥	🟥	🟥	🟥
G. max	🟥		🟥									🟥		🟥								🟥			
S. tuberosum			🟥							🟥	🟥	🟥		🟥							🟥				
V. vinifera	🟥		🟥							🟥	🟥	🟥				🟥						🟥			
T. aestivum			🟥											🟥											

Figure S10. miRNA-family distribution in eudicotyledons and monocotyledons. A total of 10 species were analyzed (*O. europaea* var. *sylvestris, B. distachyon, O. sativa, S. bicolor, Z. mays, A. thaliana, G. max, S. tuberosum, V. vinifera* and *T. aestivum*). Red: present; white: absent.

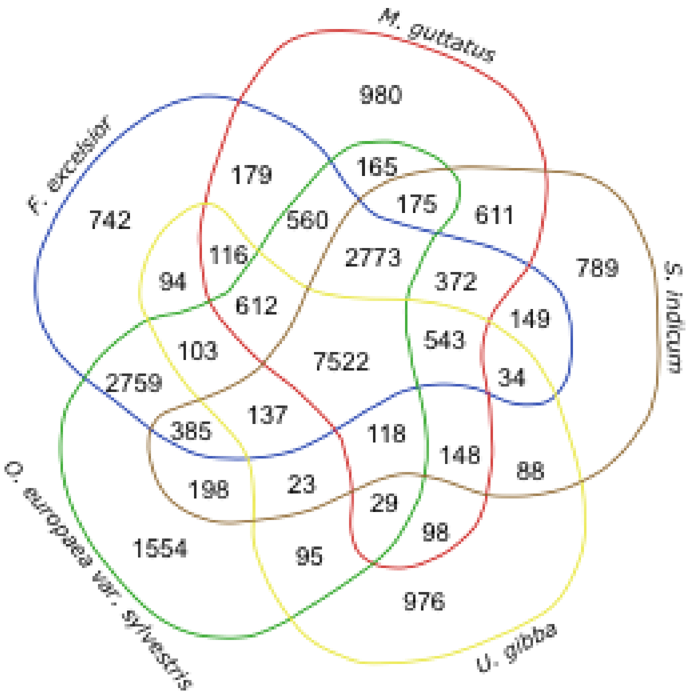


Figure S11. Comparison of gene families. Venn diagram showing gene family numbers among genomes of *S. indicum, U. gibba, M. guttatus, F. excelsior and O. europaea* var. *sylvestris.* (singleton genes not included).

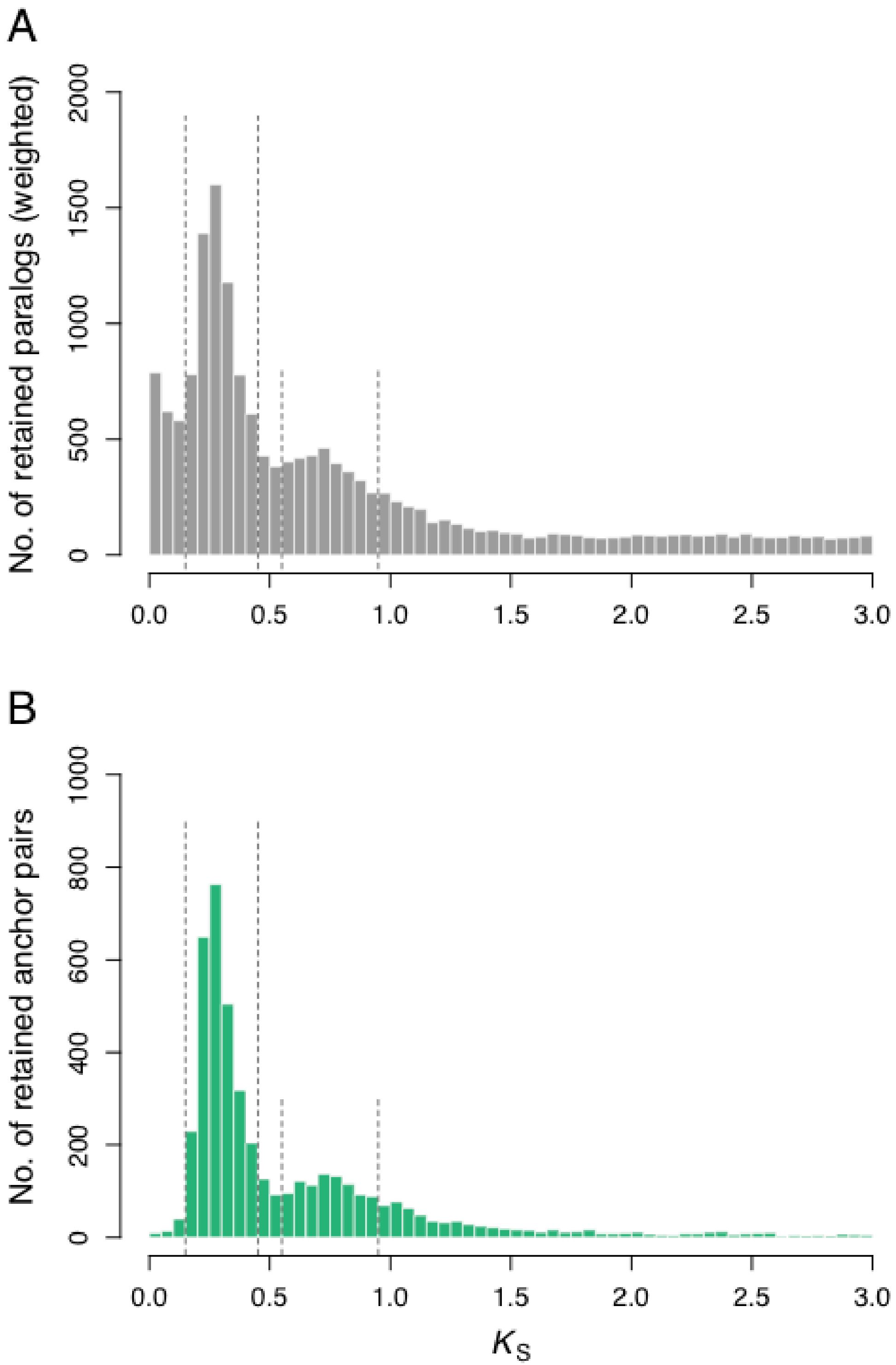


Figure S12. Oleaster ***K*****S-based age distributions. A)** Distribution of synonymous substitutions per synonymous site (*K*S) for the whole oleaster paranome. **B)** Likewise but for duplicated anchors found in collinear regions only, as identified by i-ADHoRe. Two WGD events were identified in both distributions, with peaks centered around *K*S of 0.25 and 0.75, respectively. Dashed lines indicate *K*S boundaries used to extract duplicate pairs for absolute phylogenomic dating of such WGD events.

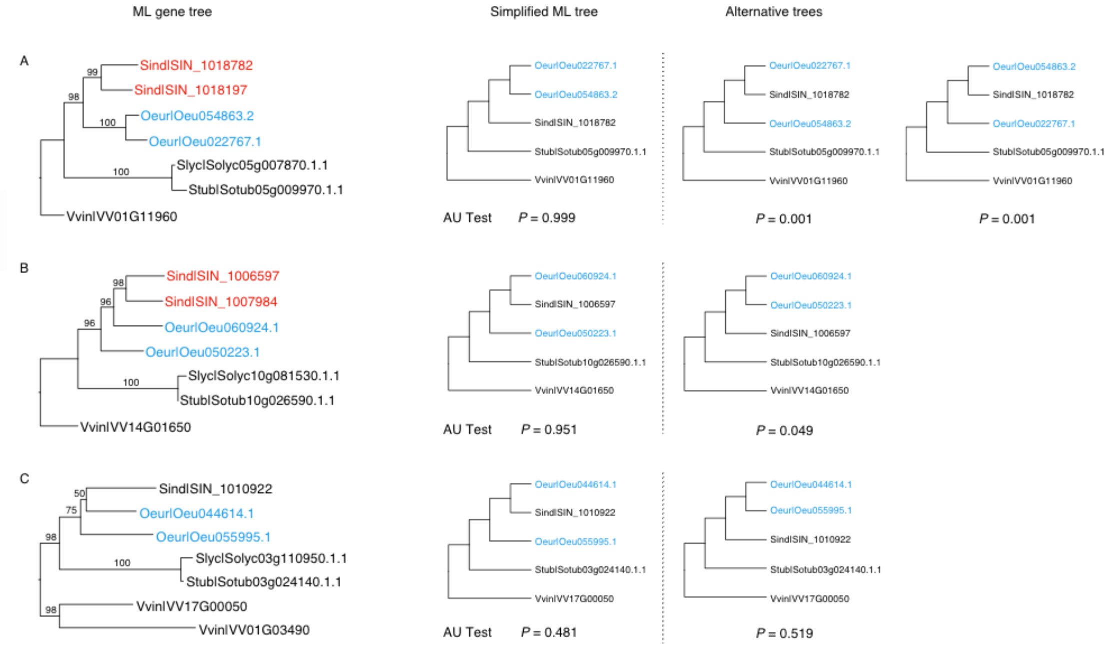


Figure S13. Examples of gene trees supporting two hypothetical scenarios for the older oleaster WGD. A) Gene tree supporting an independent WGD in the Oleaceae lineage. **B** and **C)** Gene trees supporting a WGD shared between oleaster and *S. indicum*. Genes in blue are duplicate genes from oleaster retained in collinear regions (anchor pairs) with K_S values ranging from 0.55 to 0.95. Genes in red are duplicate genes from *S. indicum* retained in collinear regions. The numbers on nodes in the gene trees are bootstrap values. Simplified ML topologies and their constructed alternative tree(s) supporting the rival WGD scenario are shown next to the original gene tree. *P*-values are from the Approximately Unbiased (AU) tests. In **A)** and **B)**, the simplified ML topology is exclusively supported (the alternative tree(s) are rejected with significance, *P*-values < 0.05), but in **C)** neither the simplified ML topology nor the alternative tree are significantly rejected by the AU tests (*P*-values < 0.05).

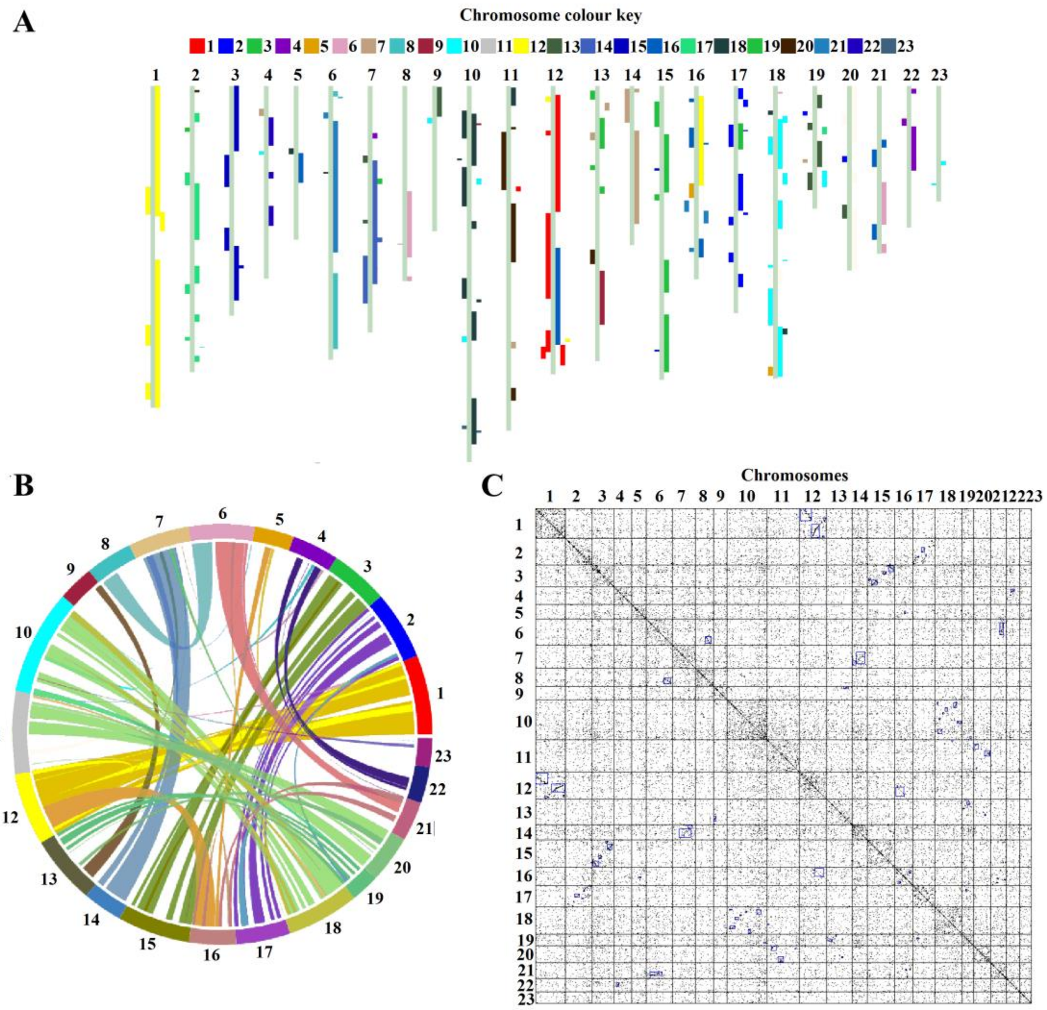


Figure S14. Microsynteny between oleaster chromosomes. A) Color representation showing homologous genomic-blocks. **B)** Inter- and intra-chromosomal syntenic relationships. **C)** Dot-plot representation of syntenic blocks among oleaster chromosomes.

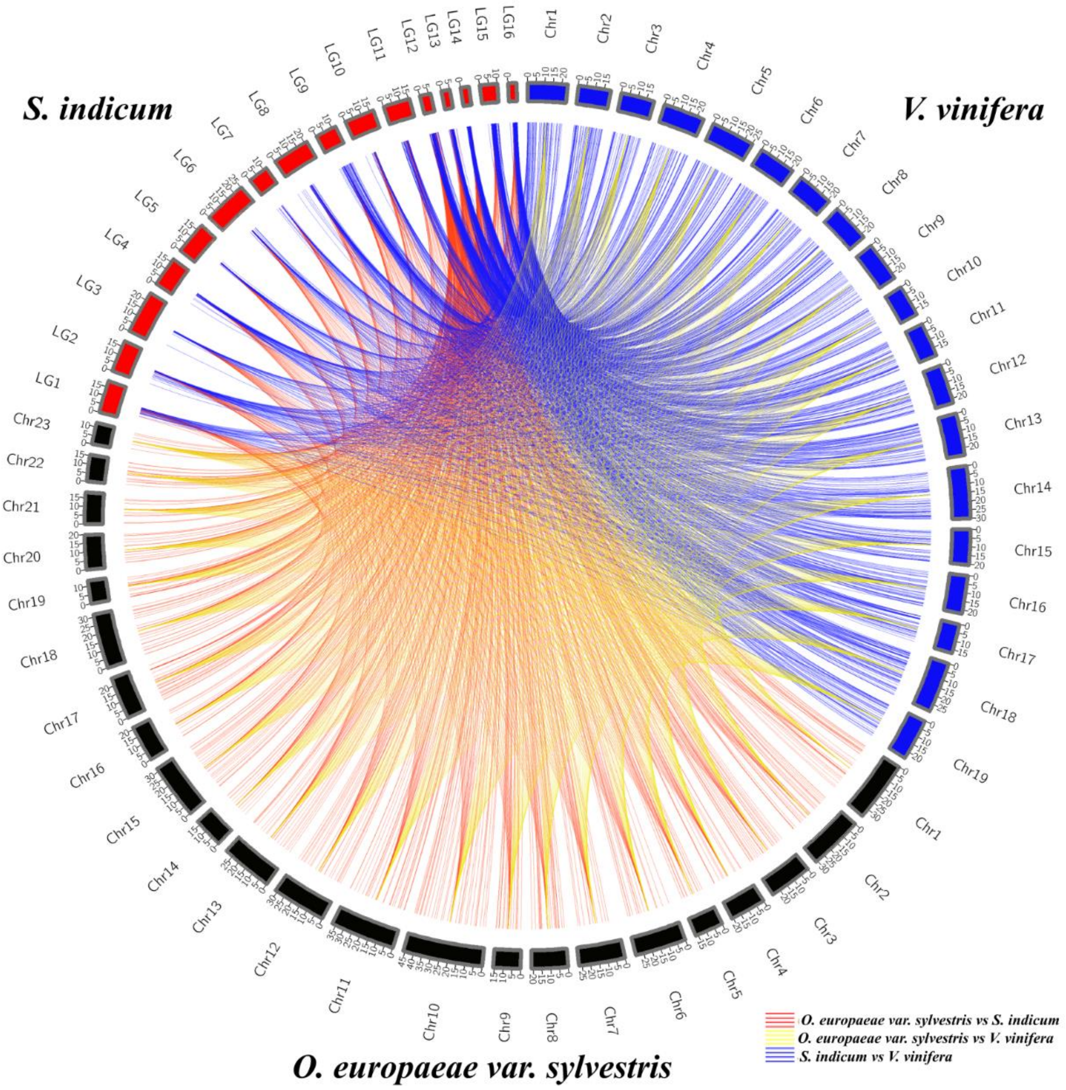


Figure S15. Synteny analyses among oleaster, sesame and grapevine. Large numbers of syntenic block were identified when comparing either the oleaster genome to the grapevine or sesame genome.

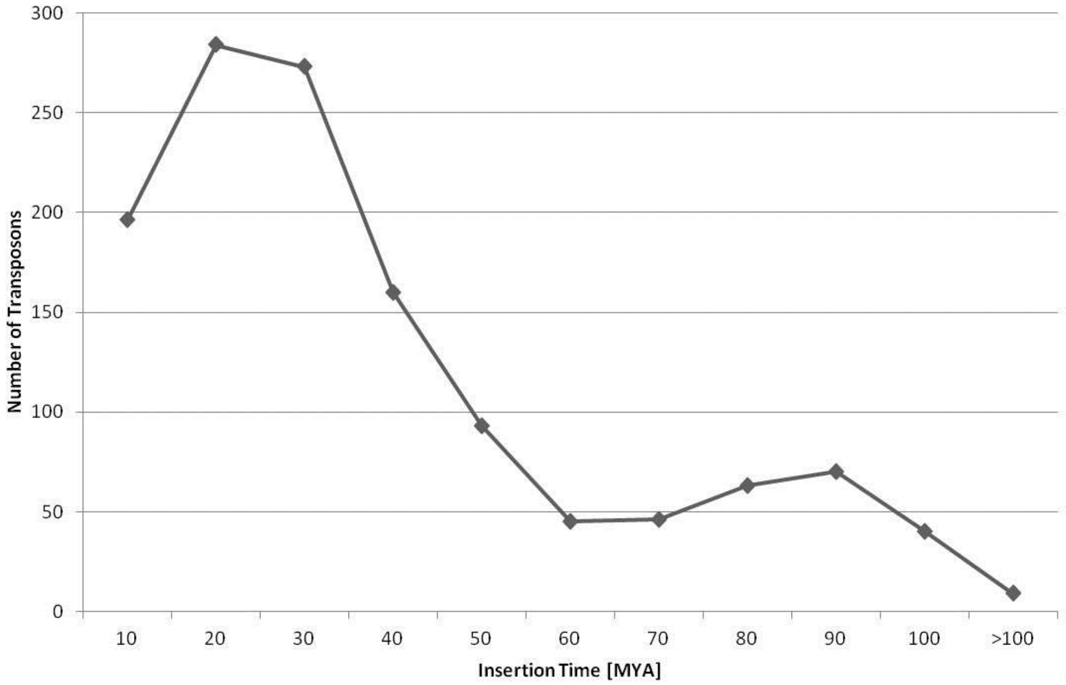


Figure S16. Oleaster full-length LTR-retrotransposon time of insertion predictions. Substitution rates between LTR were calculated as twice the mean number of synonymous substitutions per site, per year.

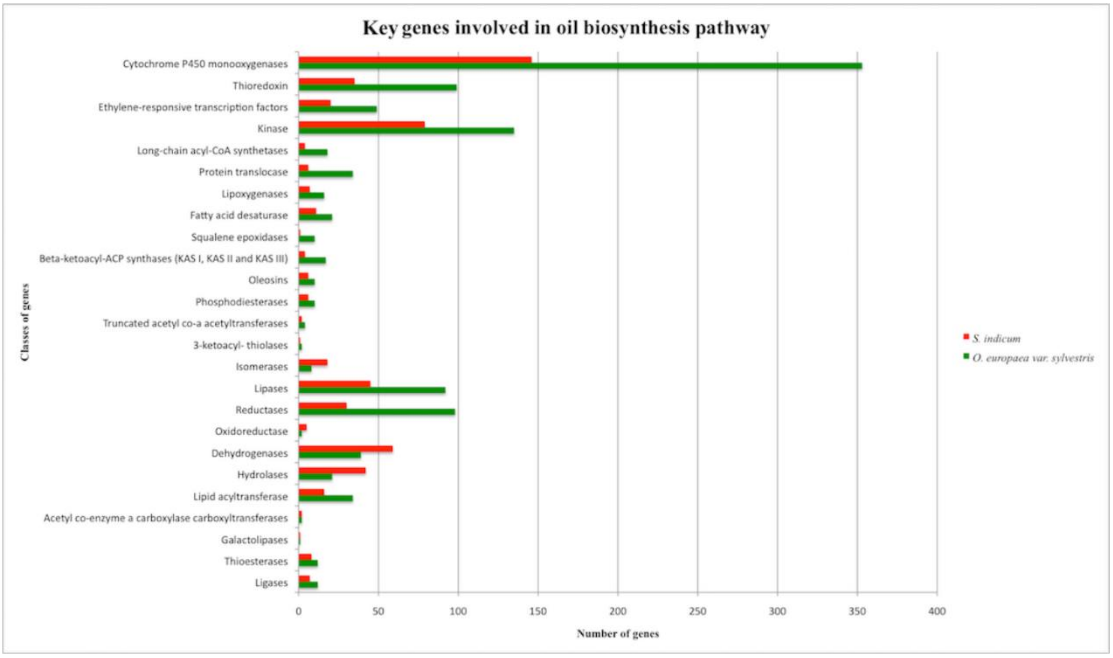


Figure S17. Key genes involved in oil biosynthesis pathway. Bar chart of genes involved in oil biosynthesis in sesame and oleaster.

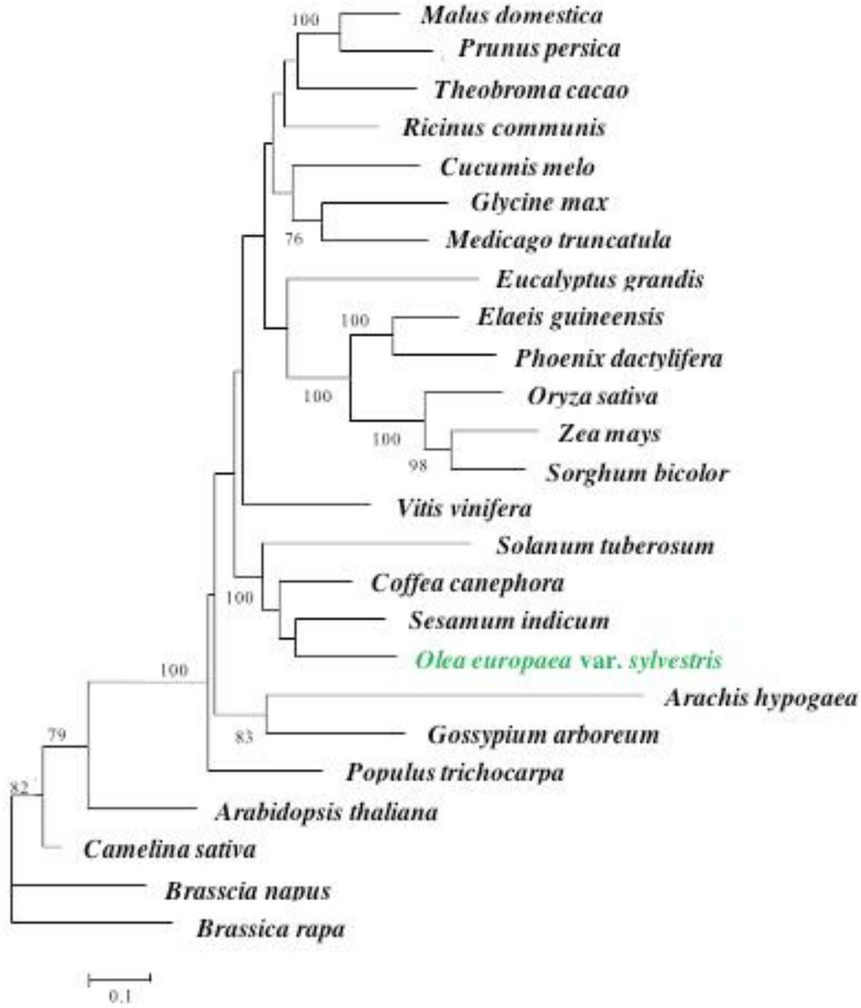


Figure S18. Phylogeny based on oil biosynthesis genes. Eight oil biosynthesis genes of selected 26 plant species were aligned to show evolutionary similarities.

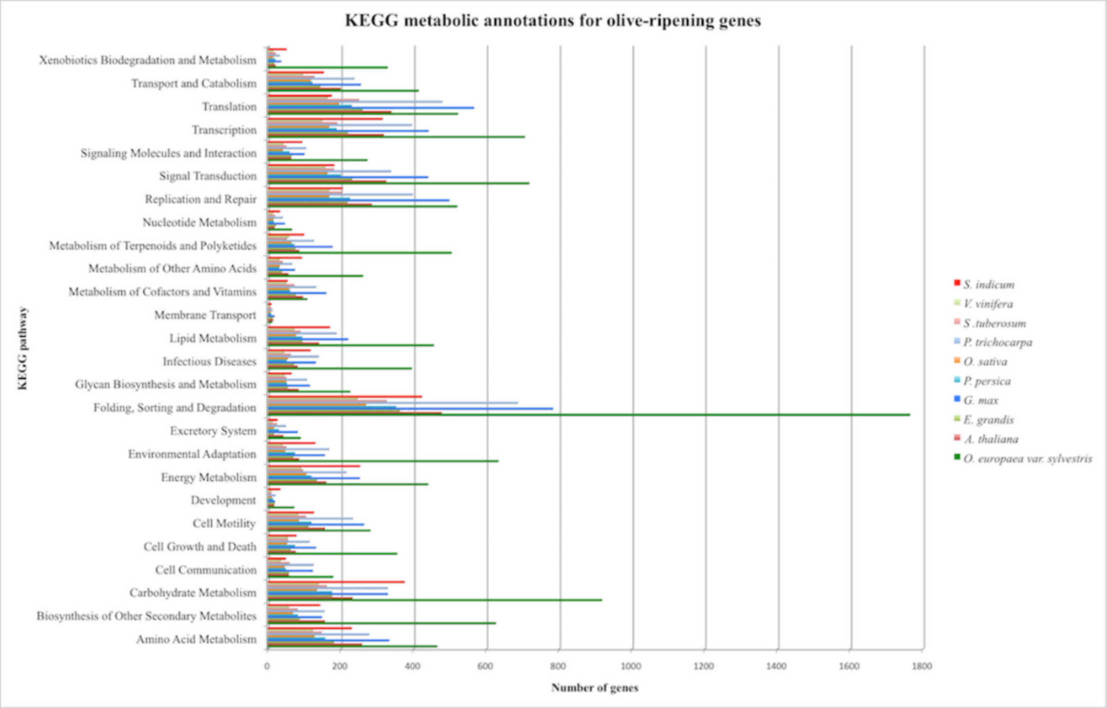


Figure S19. KEGG annotation of ripening genes. Bar chart of genes involved in ripening biosynthesis in 10 plant species: *A. thaliana, O. sativa, P. trichocarpa, S. indicum, E. grandis*, *V. vinifera*, *S. tuberosum*, *P. persica*, *G. max* and *O. europaea* var. *sylvestris*.

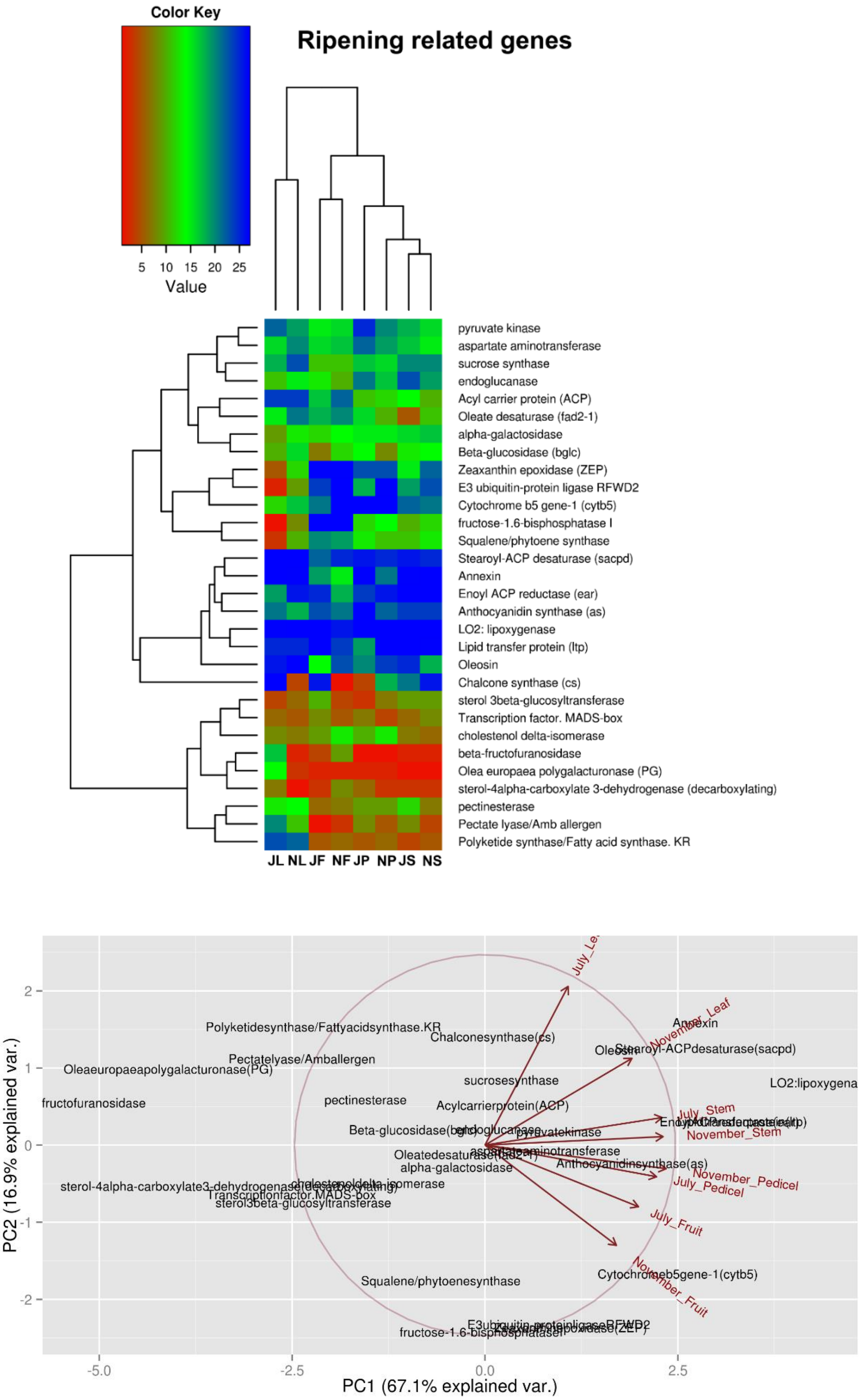
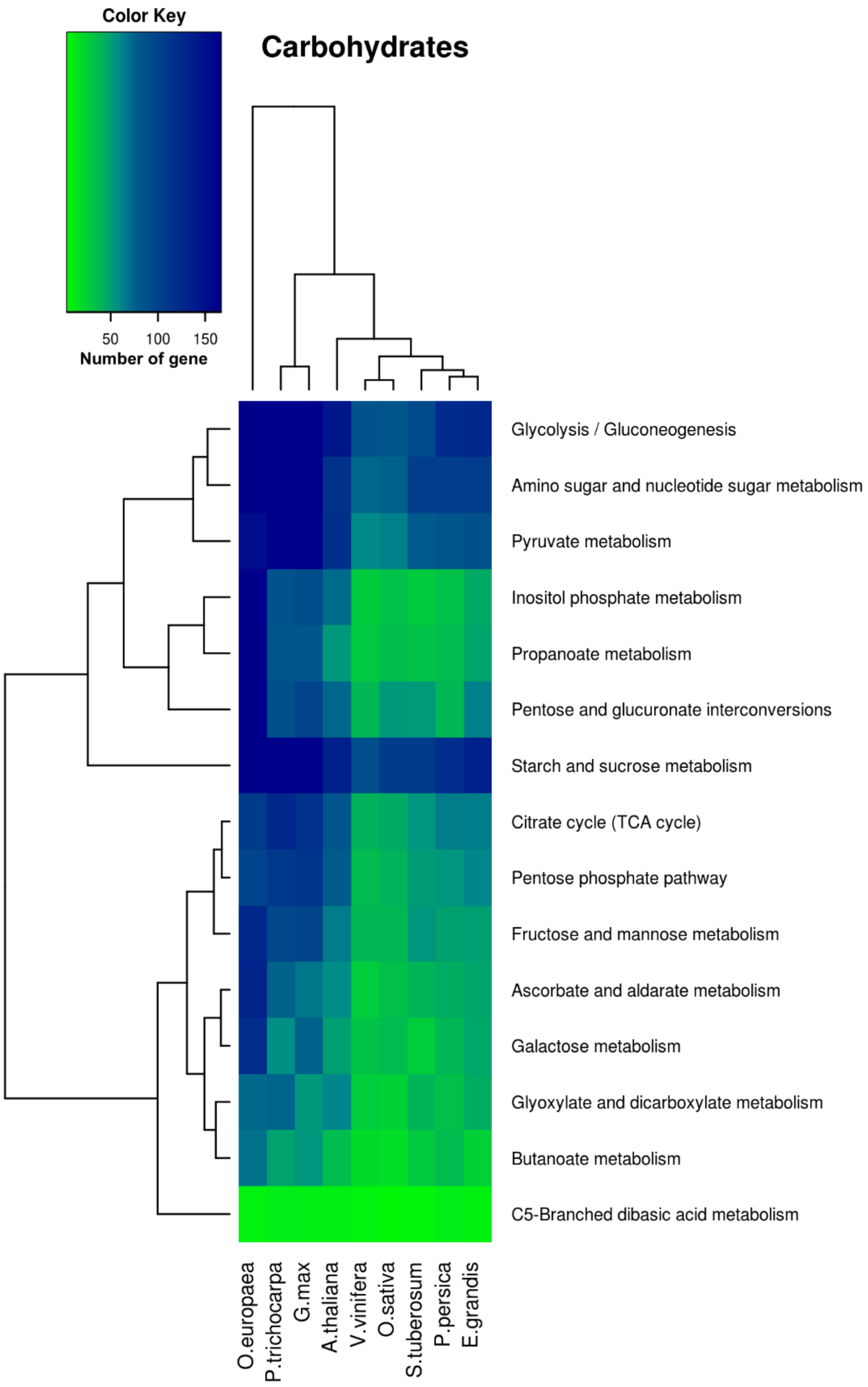


Figure S20. Expression analyses of ripening genes. A) Genes involved in ripening are shown in heatmap (log2 RPKM counts using eight RNA-seq libraries). **B)** Principal-component analysis (PCA) of gene expression, showing percentage of variations explained by two principal components.


Color Key
50
100
150
Number of gene
Carbohydrates
Glycolysis / Gluconeogenesis
Amino sugar and nucleotide sugar metabolism
Pyruvate metabolism
Inositol phosphate metabolism
Propanoate metabolism
Pentose and glucuronate interconversions
Starch and sucrose metabolism
Citrate cycle (TCA cycle)
Pentose phosphate pathway
Fructose and mannose metabolism
Ascorbate and aldarate metabolism
Galactose metabolism
Glyoxylate and dicarboxylate metabolism
Butanoate metabolism
C5-Branched dibasic acid metabolism
O.europaea
P.trichocarpa
G.max
A.thaliana
V.vinifera
O.sativa
S.tuberosum
P.persica
E.grandis

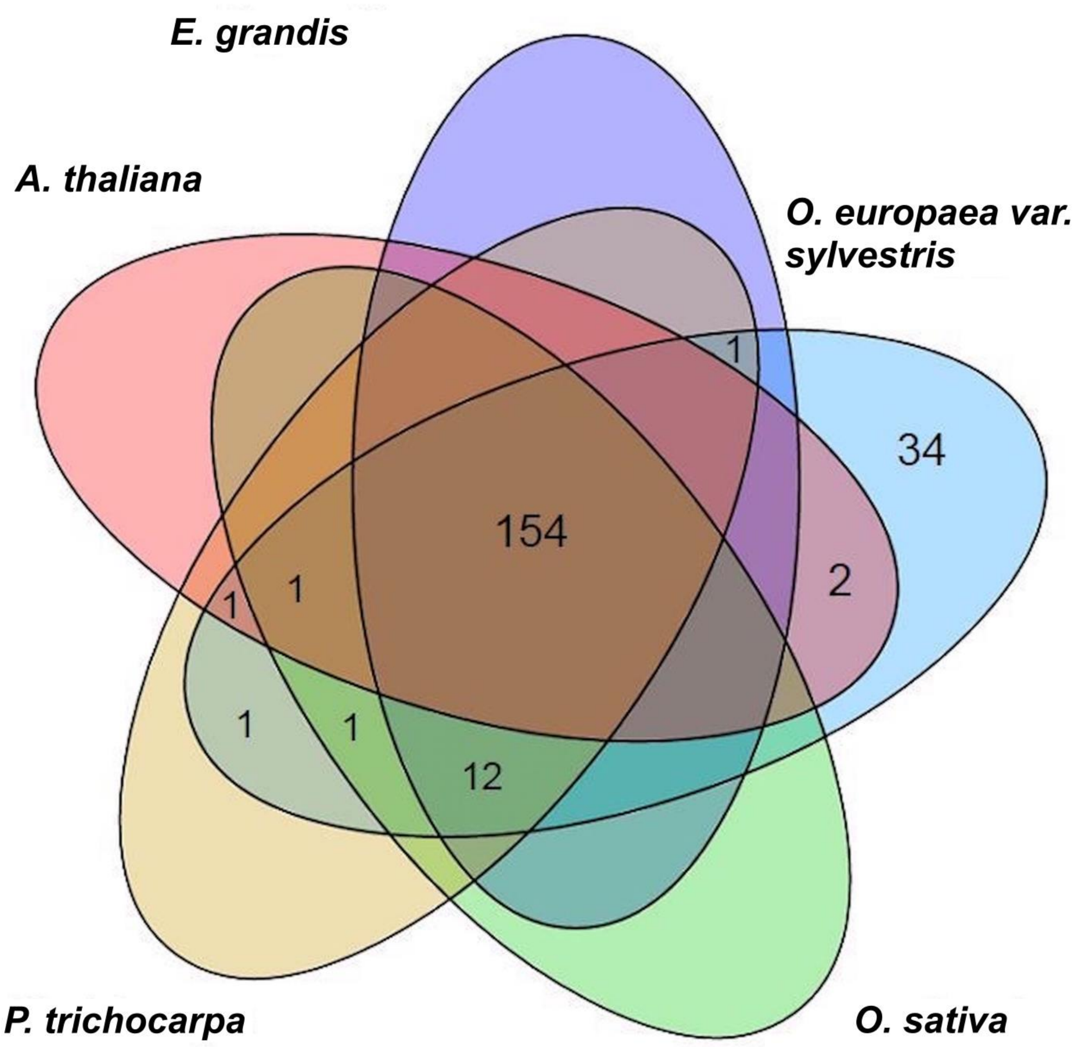
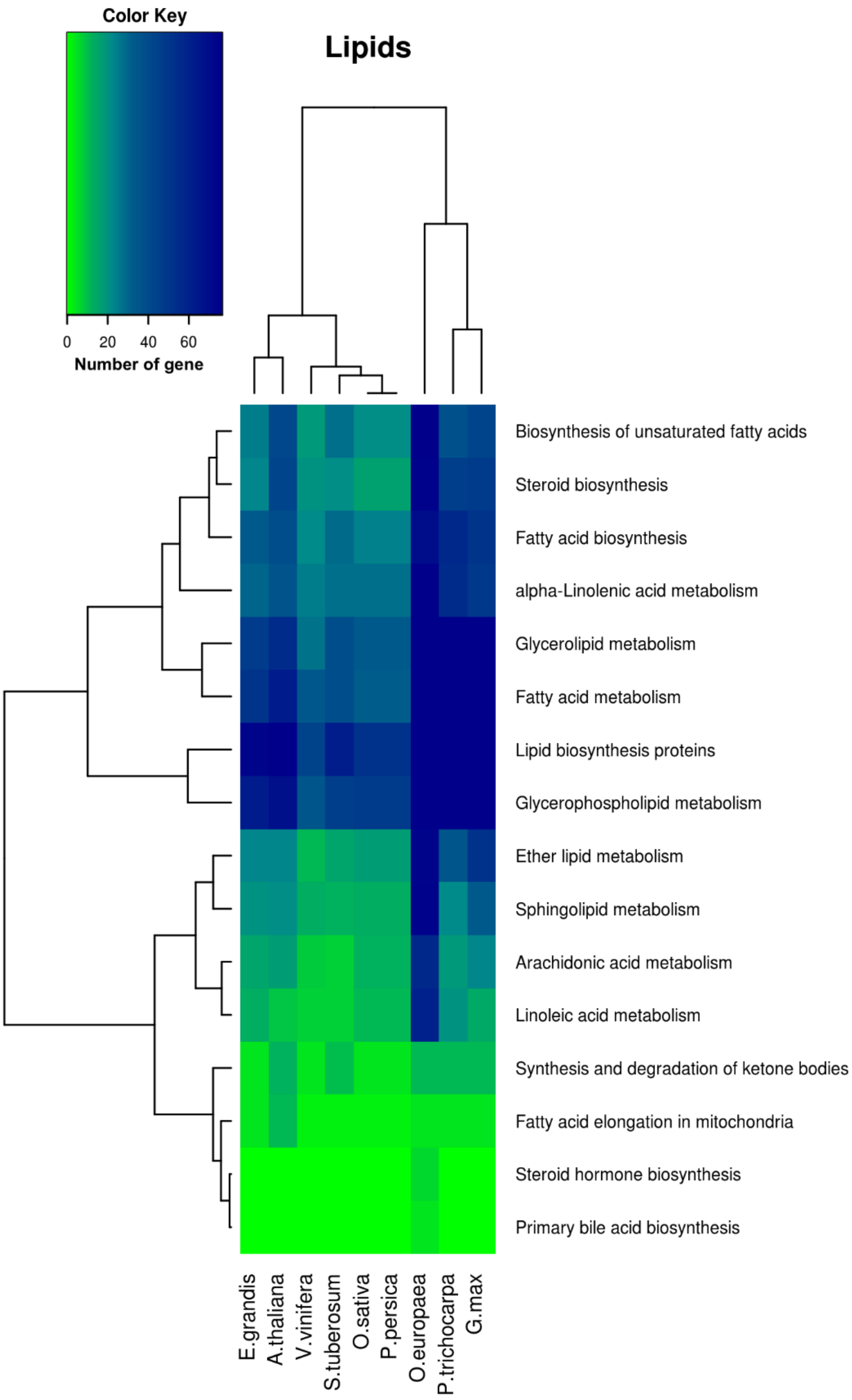


Figure S21. Carbohydrate metabolism genes of five species. A) Genes involved in carbohydrate metabolism are shown in heatmap indicating the numbers of genes present in each species. **B)** Venn diagram for genes involved in ripening with shared and unique KEGG-annotation numbers.


Color Key
Lipids
0
20
40
60
Number of gene
Biosynthesis of unsaturated fatty acids
Steroid biosynthesis
Fatty acid biosynthesis
alpha-Linolenic acid metabolism
Glycerolipid metabolism
Fatty acid metabolism
Lipid biosynthesis proteins
Glycerophospholipid metabolism
Ether lipid metabolism
Sphingolipid metabolism
Arachidonic acid metabolism
Linoleic acid metabolism
Synthesis and degradation of ketone bodies
Fatty acid elongation in mitochondria
Steroid hormone biosynthesis
Primary bile acid biosynthesis
E.grandis
A.thaliana
V.vinifera
S.tuberosum
O.sativa
P.persica
O.europaea
P.trichocarpa
G.max

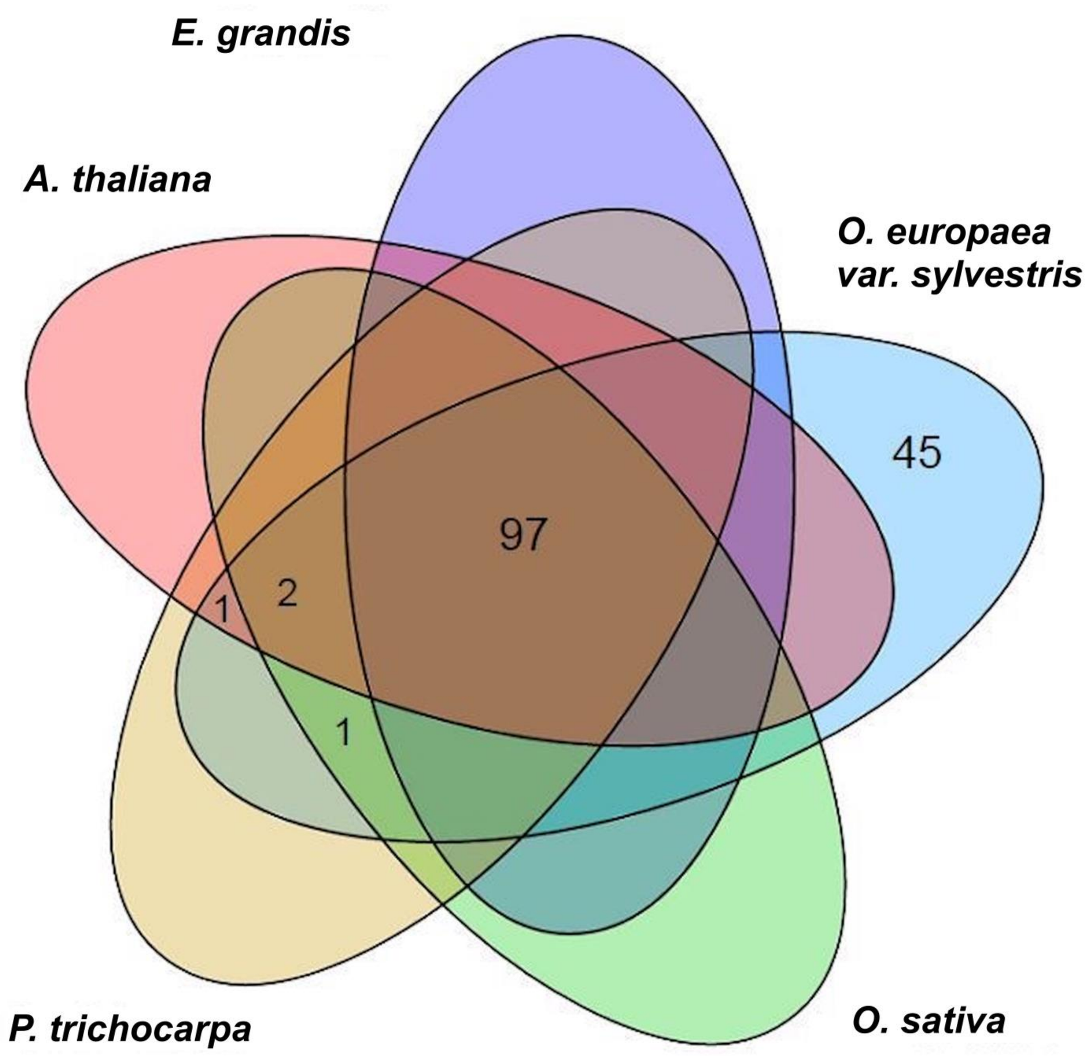
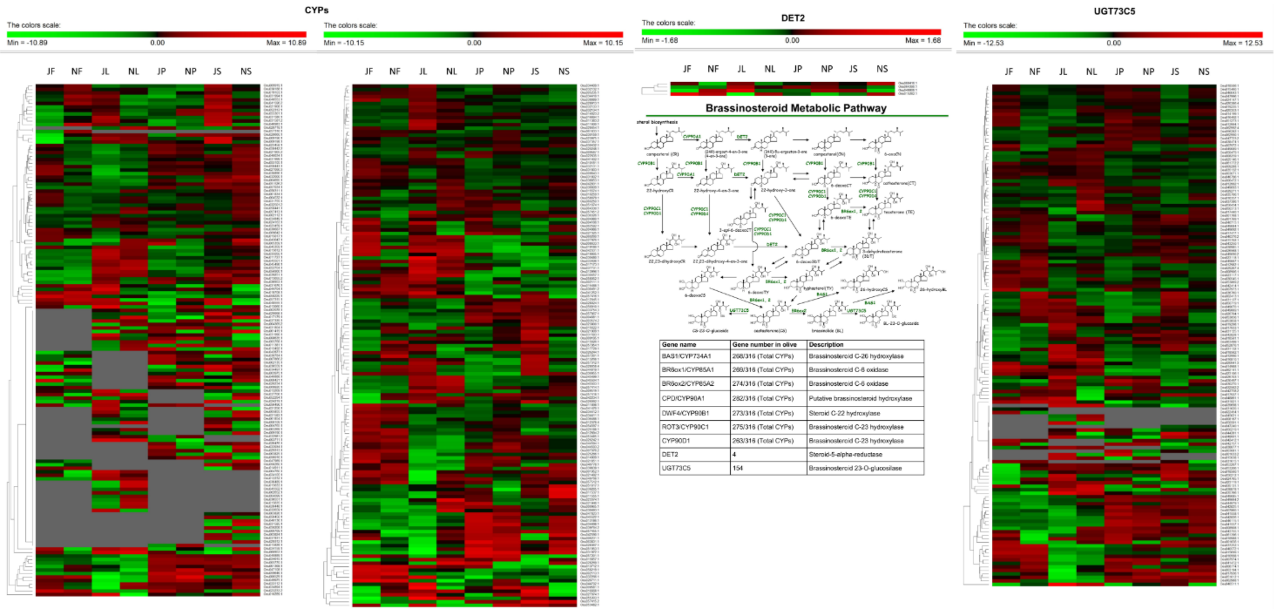
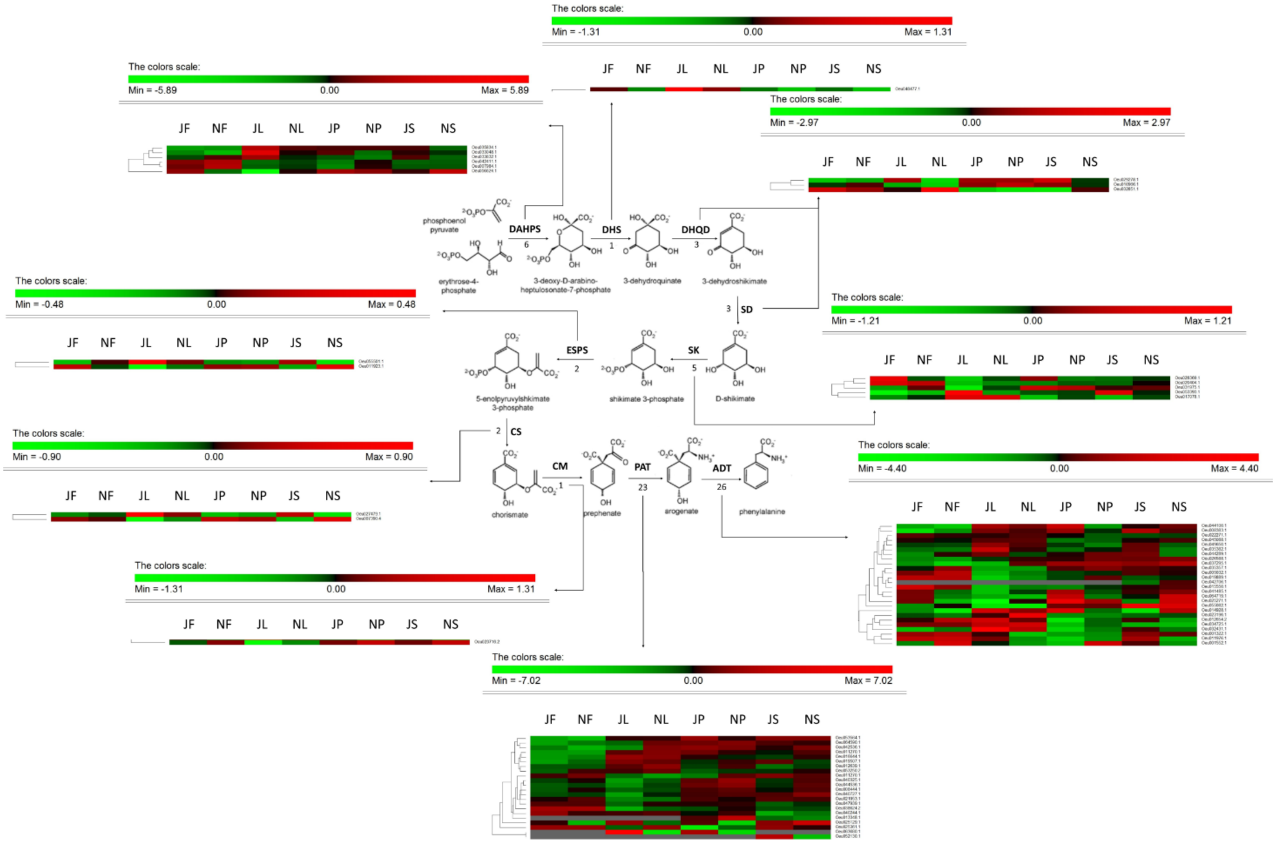
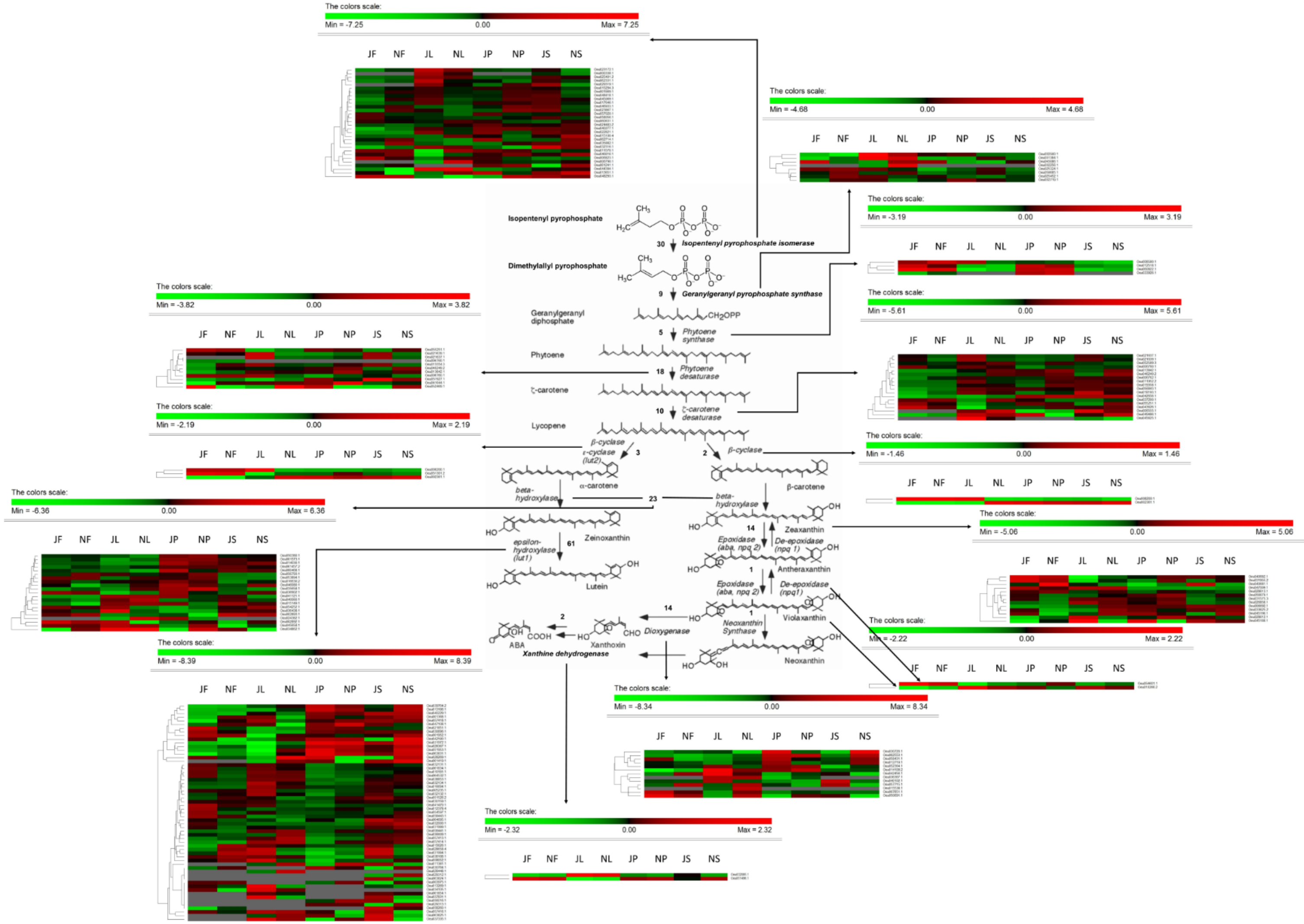


Figure S22. Lipid metabolism genes of five species. A) Genes involved in lipid metabolism are shown in heatmap indicating the numbers of genes present in each species **B)** Venn diagram for such genes, with shared and unique KEGG-annotation numbers.



Gene name	Gene number in olive	Description
BAS1/CYP734A1	268/316 (Total CYPs)	Brassinosteroid C-26 hydroxylase
BR6ox1/CYP85A1	269/316 (Total CYPs)	Brassinosteroid C-6 oxidase
BR6ox2/CYP85A2	274/316 (Total CYPs)	Brassinosteroid C-6 oxidase
CPD/CYP90A1	282/316 (Total CYPs)	Putative brassinosteroid hydroxylase
DWF4/CYP90B1	273/316 (Total CYPs)	Steroid C-22 hydroxylase
ROT3/CYP90C1	275/316 (Total CYPs)	Brassinosteroid C-23 hydroxylase
CYP90D1	263/316 (Total CYPs)	Brassinosteroid C-23 hydroxylase
DET2	4	Steroid-5-alpha-reductase
UGT73C5	154	Brassinosteroid 23-O-glucosilase




The colors scale:
Min = -7.25
0.00
Max = 7.25
JF NF JL NL JP NP JS NS
The colors scale:
Min = -4.68
0.00
Max = 4.68
The colors scale:
Min = -3.19
0.00
Max = 3.19
The colors scale:
Min = -5.61
0.00
Max = 5.61
The colors scale:
Min = -1.46
0.00
Max = 1.46
The colors scale:
Min = -5.06
0.00
Max = 5.06
The colors scale:
Min = -2.22
0.00
Max = 2.22
The colors scale:
Min = -8.34
0.00
Max = 8.34
The colors scale:
Min = -2.32
0.00
Max = 2.32
The colors scale:
Min = -8.39
0.00
Max = 8.39
The colors scale:
Min = -6.36
0.00
Max = 6.36
The colors scale:
Min = -2.19
0.00
Max = 2.19
The colors scale:
Min = -3.82
0.00
Max = 3.82
Isopentenyl pyrophosphate
30 Isopentenyl pyrophosphate isomerase
Dimethylallyl pyrophosphate
9 Geranylgeranyl pyrophosphate synthase
Geranylgeranyl diphosphate
CH2OPP
5 Phytoene synthase
Phytoene
18 Phytoene desaturase
ζ-carotene
10 ζ-carotene desaturase
Lycopene
β-cyclase
ε-cyclase (lut2)
3
2 β-cyclase
α-carotene
β-carotene
beta-hydroxylase
23
beta-hydroxylase
Zeinoxanthin
61
epsilon-hydroxylase (lut1)
Lutein
14
Zeaxanthin
Epoxidase (aba, npq 2)
De-epoxidase (npq 1)
Antheraxanthin
1
Epoxidase (aba, npq 2)
De-epoxidase (npq1)
Violaxanthin
1
Neoxanthin Synthase
Neoxanthin
14
Dioxygenase
Xanthoxin
2
ABA
COOH
CHO
Xanthine dehydrogenase
HO
OH

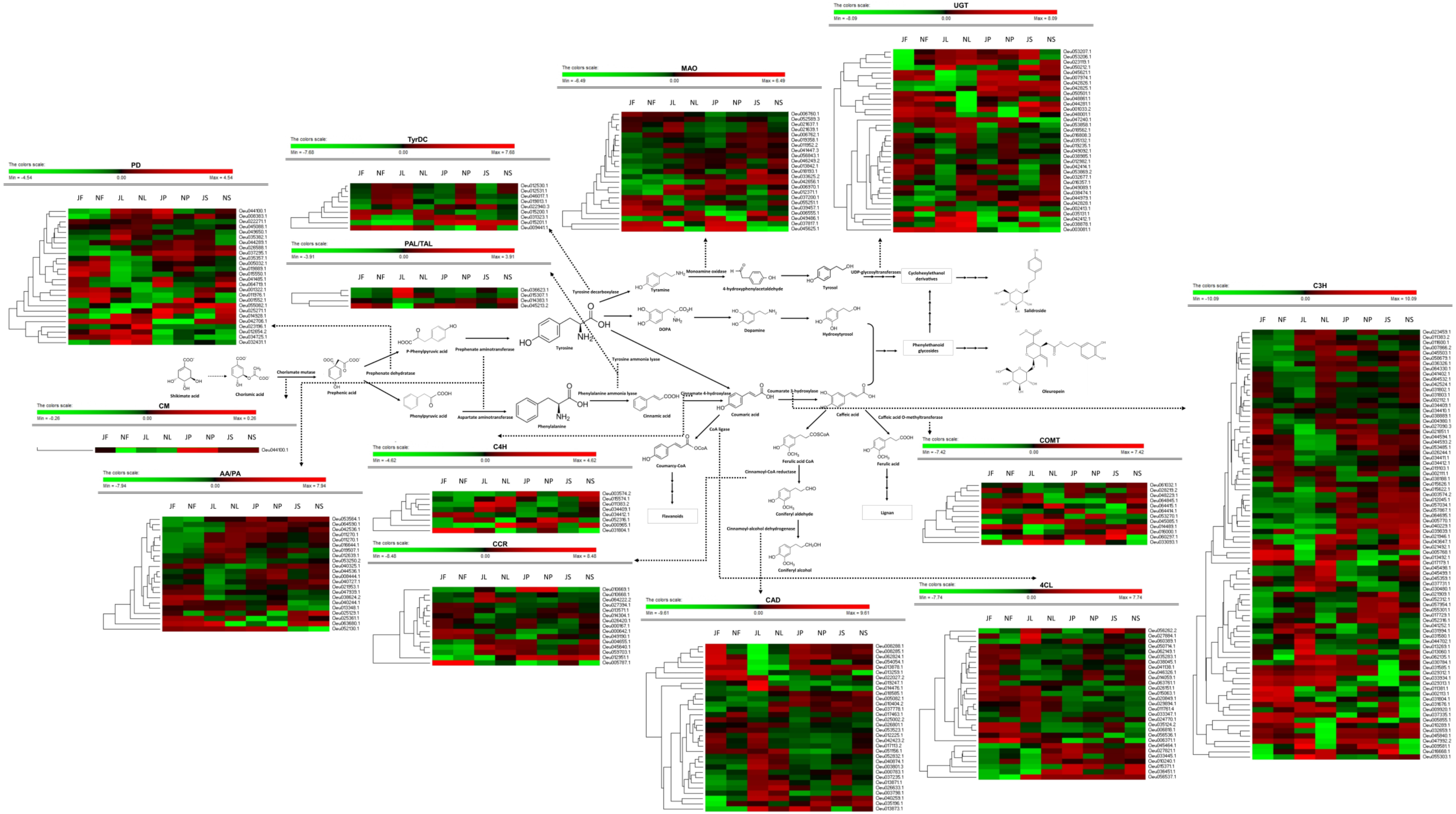
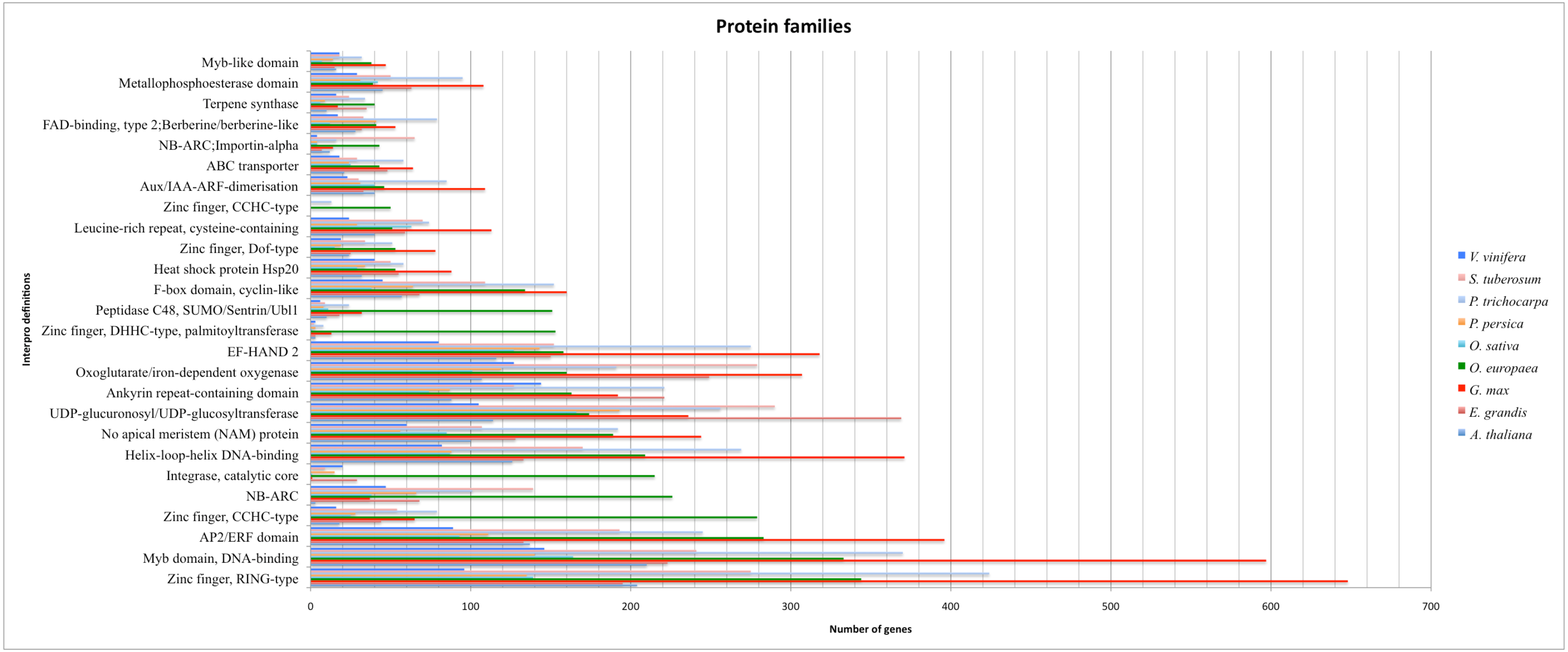

PD
The colors scale:
Min = -4.54
0.00
Max = 4.54
JF NF JL NL JP NP JS NS
TyrDC
Min = -7.68
Max = 7.68
PAL/TAL
Min = -3.91
Max = 3.91
MAO
Min = -6.49
Max = 6.49
UGT
Min = -8.09
Max = 8.09
C3H
Min = -10.09
Max = 10.09
CM
Min = -0.26
Max = 0.26
AA/PA
Min = -7.94
Max = 7.94
C4H
Min = -4.62
Max = 4.62
CCR
Min = -8.48
Max = 8.48
CAD
Min = -9.61
Max = 9.61
COMT
Min = -7.42
Max = 7.42
4CL
Min = -7.74
Max = 7.74
Shikimate acid
Chorismic acid
Chorismate mutase
Prephenic acid
Prephenate dehydratase
P-Phenylpyruvic acid
Prephenate aminotransferase
Tyrosine
Phenylpyruvic acid
Aspartate aminotransferase
Phenylalanine
Phenylalanine ammonia lyase
Cinnamic acid
Cinnamate 4-hydroxylase
Coumaric acid
Tyrosine ammonia lyase
Tyrosine decarboxylase
Tyramine
Monoamine oxidase
4-hydroxyphenylacetaldehyde
Tyrosol
UDP-glycosyltransferases
Cyclohexylethanol derivatives
Salidroside
DOPA
Dopamine
Hydroxytyrosol
Phenylethanoid glycosides
Oleuropein
Coumarate 3-hydroxylase
Caffeic acid
Caffeic acid O-methyltransferase
Ferulic acid
Lignan
CoA ligase
Coumaroyl-CoA
Flavanoids
Ferulic acid CoA
Cinnamoyl-CoA reductase
Coniferyl aldehyde
Cinnamoyl-alcohol dehydrogenase
Coniferyl alcohol

Figure S23. Secondary metabolite biosynthesis pathways in oleaster. Genes involved in **A)** Brassinosteroid and **B)** Shikimate pathways. **C)** Carotenoid pathways are represented with arrows, and their expression quantifications are exhibited by heatmaps. **D)** Oleuropein biosynthesis pathway. Phenylethanoid (PE) pathway for oleuropein biosynthesis is depicted and the genes involved with their expression patterns are represented. The pathway starts with shikimate acid and continues with the production of two main amino acids, tyrosine and phenylalanine. Through enzymatic reactions from these amino acids, oleuropein is produced. Gene name abbreviations are as follows: AA: Aspartate aminotransferase, CM: Chorismate mutase, COMT: Caffeic acid O-methyltransferase, C4H: Cinnamate-4-hydroxylase, CCR: Cinnamoyl-CoA reductase, CAD: cinnamyl alcohol dehydrogenase, 4CL: Coenzyme A ligase, C3H: coumarate 3-hydroxylase, MAO: Monoamine oxidase, PAL: Phenylalanine ammonia lyase, TAL: Tyrosine ammonia lyase, PD: Prephenate dehydratase, TyrDC: Tyrosine decarboxylase, UGT: UDP glycosyltransferase, J: July, N: November, L: Leaf, S: Stem, P: Pedicel, F: Fruit.


Protein families
Interpro definitions
Myb-like domain
Metallophosphoesterase domain
Terpene synthase
FAD-binding, type 2;Berberine/berberine-like
NB-ARC;Importin-alpha
ABC transporter
Aux/IAA-ARF-dimerisation
Zinc finger, CCHC-type
Leucine-rich repeat, cysteine-containing
Zinc finger, Dof-type
Heat shock protein Hsp20
F-box domain, cyclin-like
Peptidase C48, SUMO/Sentrin/Ubl1
Zinc finger, DHHC-type, palmitoyltransferase
EF-HAND 2
Oxoglutarate/iron-dependent oxygenase
Ankyrin repeat-containing domain
UDP-glucuronosyl/UDP-glucosyltransferase
No apical meristem (NAM) protein
Helix-loop-helix DNA-binding
Integrase, catalytic core
NB-ARC
Zinc finger, CCHC-type
AP2/ERF domain
Myb domain, DNA-binding
Zinc finger, RING-type
0
100
200
300
400
500
600
700
Number of genes
V. vinifera
S. tuberosum
P. trichocarpa
P. persica
O. sativa
O. europaea
G. max
E. grandis
A. thaliana

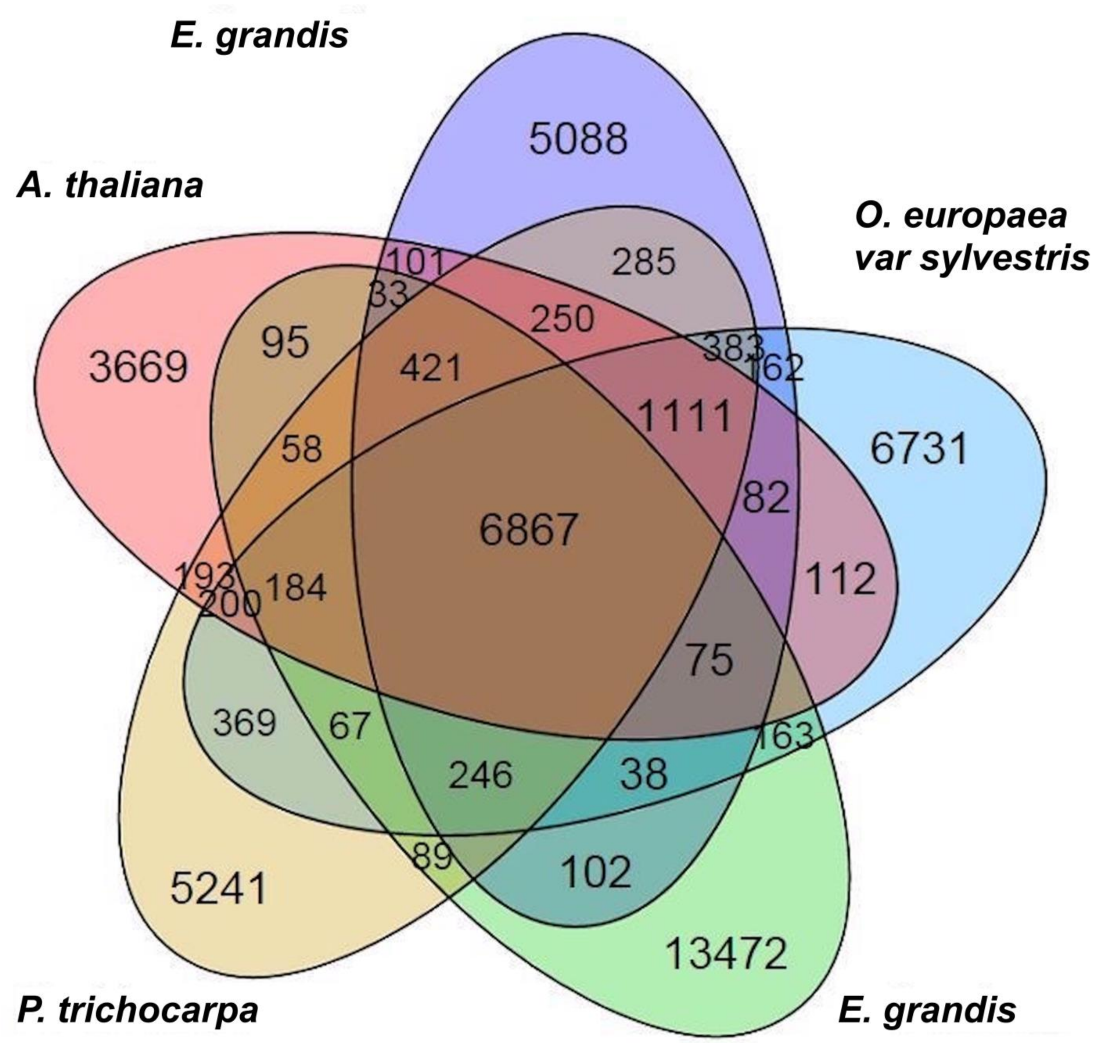


Figure S24. Comparison of protein families. A) Oleaster and the eight other plant species protein families were compared based on interpro definitions. **B)** Venn diagram based on clustering ID and InterPro definitions, for oleaster protein families compared to eight other plant species.

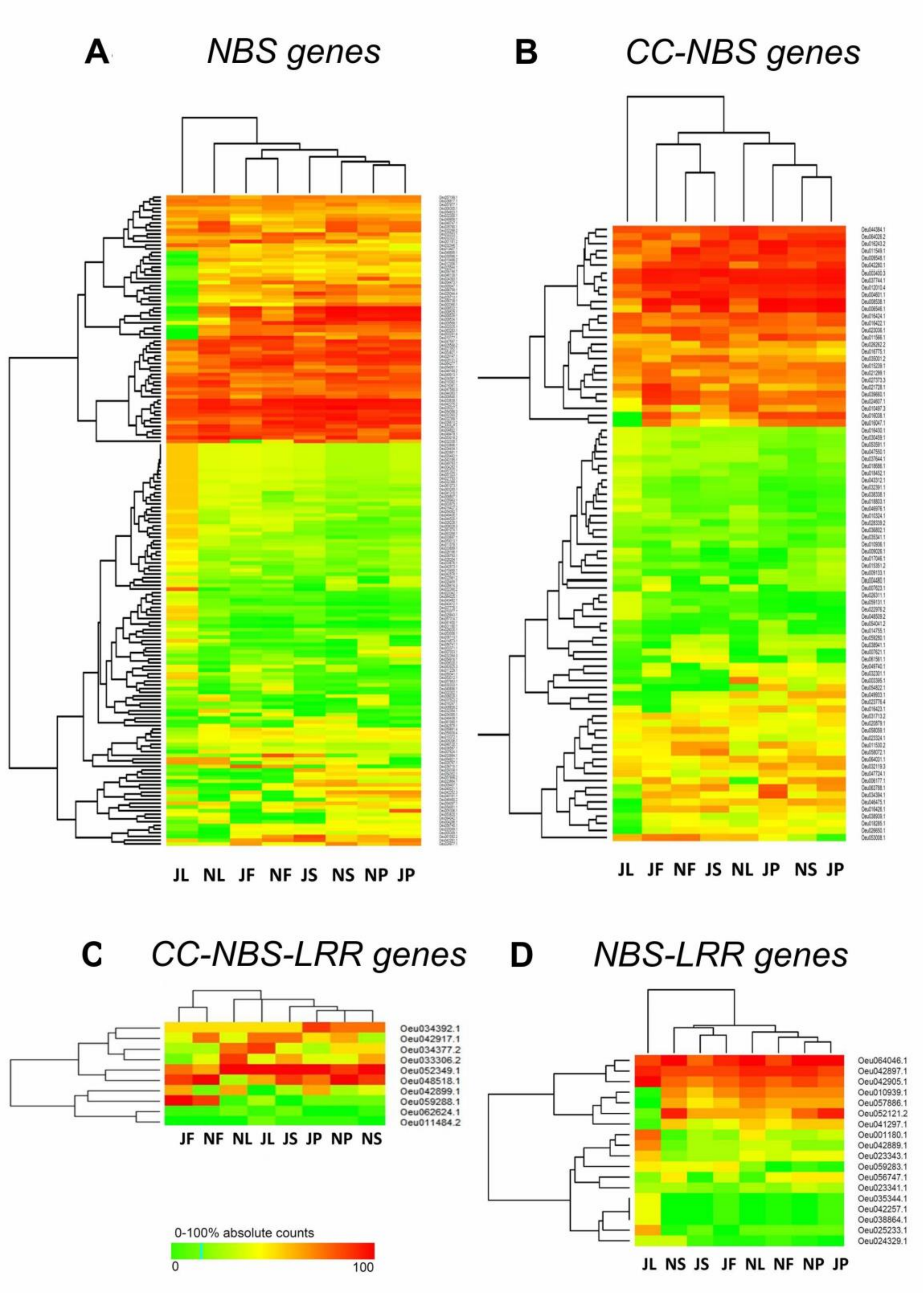


Figure S25. Disease-resistance genes of oleaster. Heatmap analyses of NBS-coding disease-resistance gene families in eight samples of oleaster. Gene families were subdivided into four groups, according to motif analyses: **A)** *NBS*, **B)** *CC-NBS*, **C)** *CC-NBS-LRR* and **D)** *NBS-LRR*.

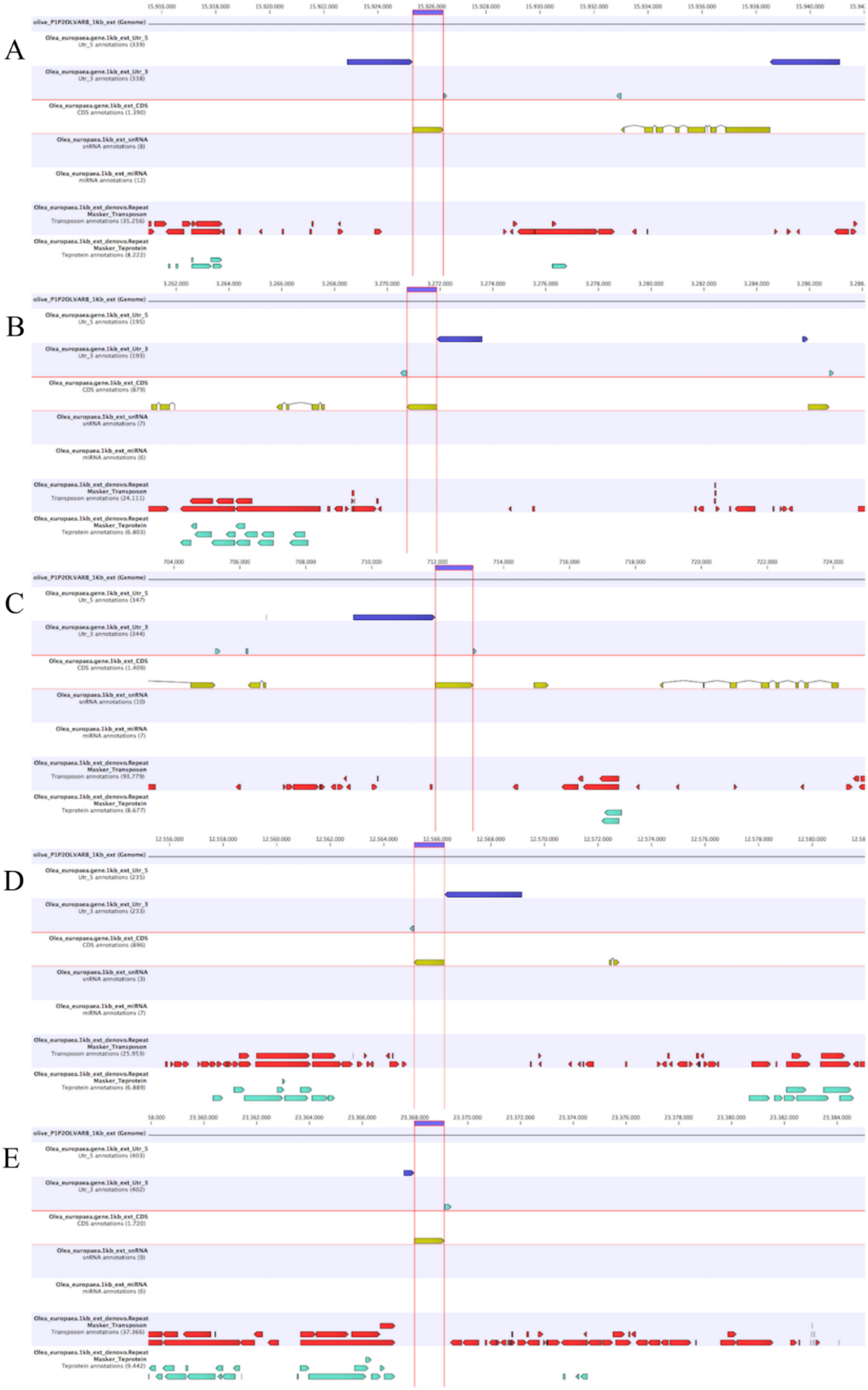
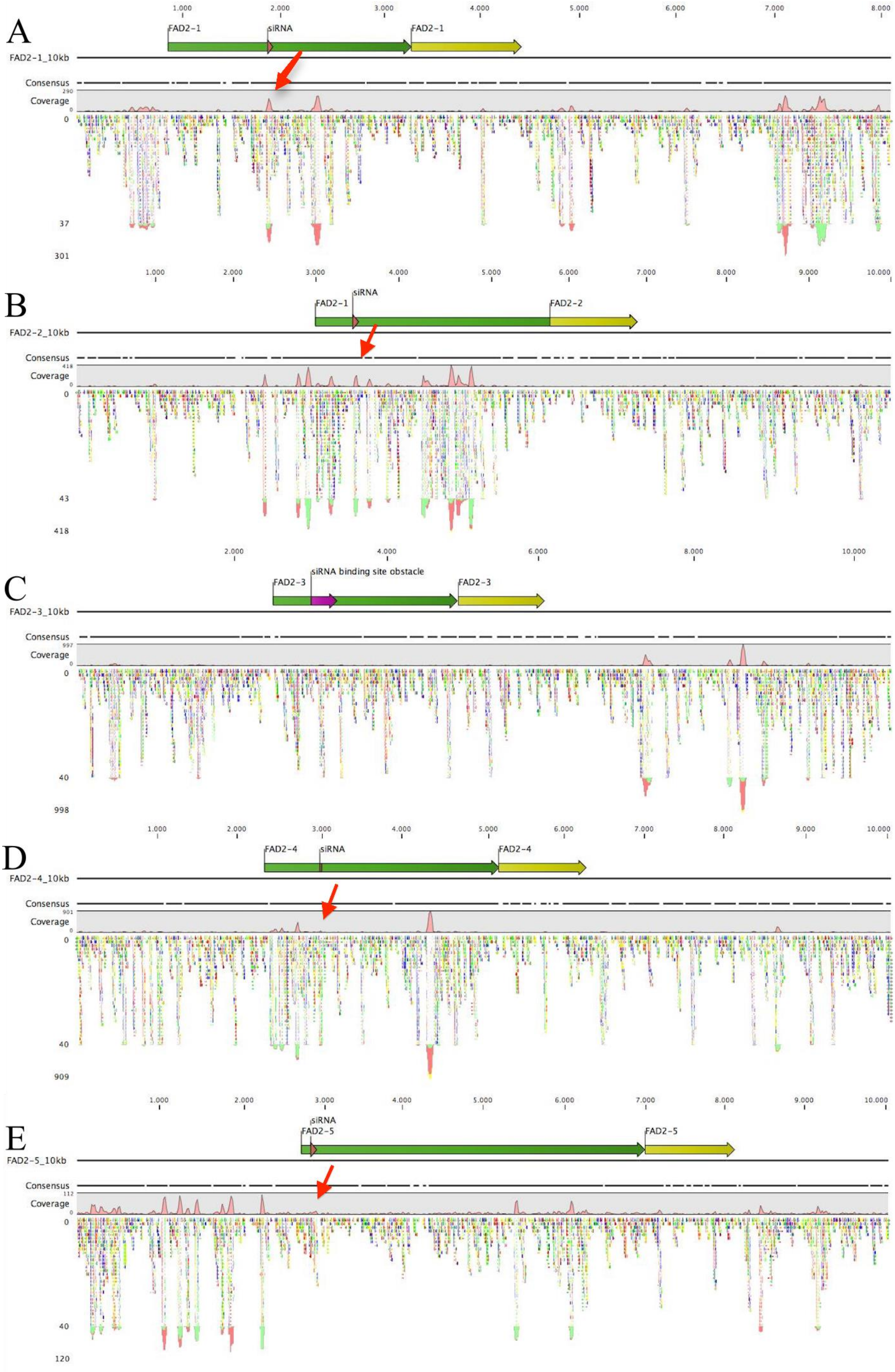


Figure S26. Representation of *FAD2*-gene genomic regions. *FAD2* genes and other genetic elements are represented for **A)** FAD2-1, **B)** FAD2-2, **C)** FAD2-3, **D)** FAD2-4 and **E)** FAD2-5. Yellow: CDS; Blue: 5'-UTR; Turquoise: 3'-UTR; Turquiosa: known transposable element; and RE: *de novo* transposon.


A
FAD2-1_10kb
FAD2-1
siRNA
FAD2-1
Consensus
Coverage
290
0
37
301
B
FAD2-2_10kb
FAD2-1
siRNA
FAD2-2
Consensus
Coverage
418
0
43
418
C
FAD2-3_10kb
FAD2-3
siRNA binding site obstacle
FAD2-3
Consensus
Coverage
997
0
40
998
D
FAD2-4_10kb
FAD2-4
siRNA
FAD2-4
Consensus
Coverage
901
0
40
909
E
FAD2-5_10kb
siRNA
FAD2-5
FAD2-5
Consensus
Coverage
112
0
40
120
1.000
2.000
3.000
4.000
5.000
6.000
7.000
8.000
9.000
10.000

F

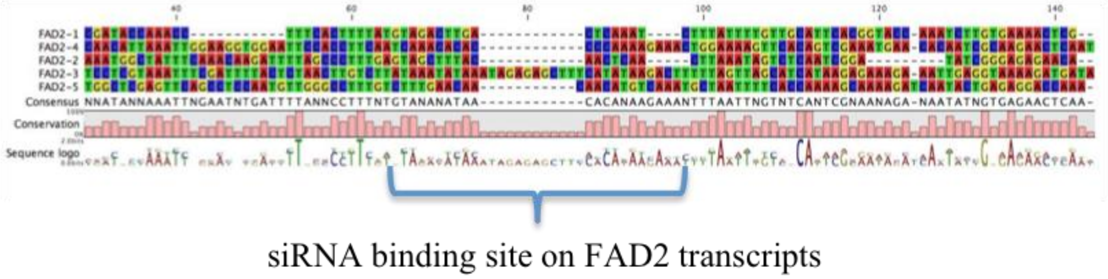
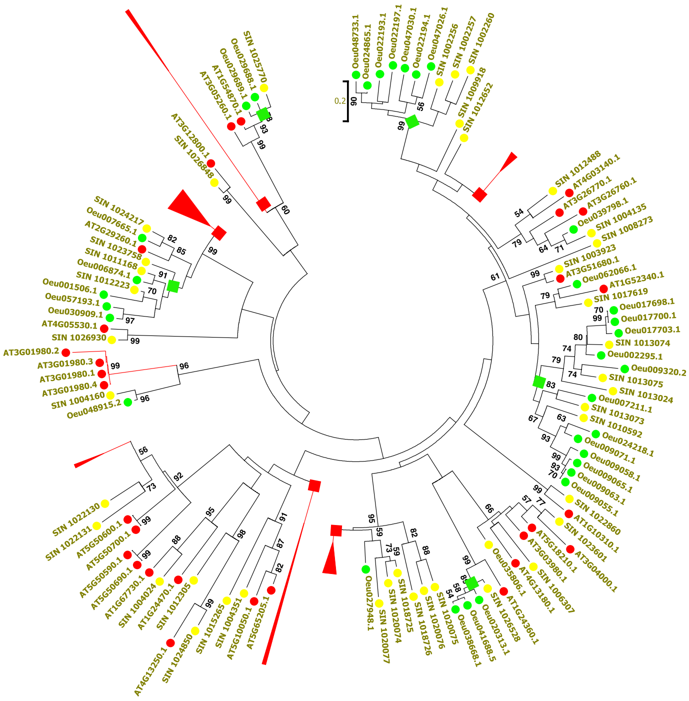
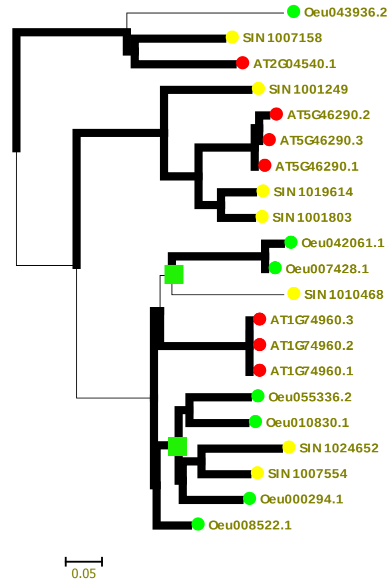

40
60
80
100
120
140
FAD2-1
FAD2-4
FAD2-2
FAD2-3
FAD2-5
Consensus
NNATANNAAATTNGAATNTGATTTTANNCCTTTNTGTANANATAA-----------CACANAAGAAANTTTAATTNGTNTCANTCGNAANAGA-NAATATNGTGAGAACTCAA-
Conservation
Sequence logo
siRNA binding site on FAD2 transcripts

Figure S27. Mapping of small-RNA reads on oleaster *FAD2* genes. Reads were taken from ripe fruit samples collected in November season. A total of 10 kbp regions including CDS and 5’-UTR of *FAD2* genes and siRNA binding-sites are represented. **A)** Mapping of sRNA reads on 10 kbp region, covering FAD2-1, **B)** FAD2-2, **C)** FAD2-3, **D)** FAD2-4 and **E)** FAD2-5. **F)** Alignment of *FAD2*-transcript siRNA binding-sites. Arrows show sRNA mapping peaks on siRNA binding-sites. Yellow regions represent CDS, green regions are 5’-UTR of *FAD2* genes and purple triangles show siRNA binding-sites. Peaks indicate sRNA maps.


0.2
Oeu048733.1
Oeu024865.1
Oeu022193.1
Oeu022197.1
Oeu047030.1
Oeu022194.1
Oeu047026.1
SIN 1002256
SIN 1002257
SIN 1002260
SIN 1009918
SIN 1012652
SIN 1012488
AT4G03140.1
AT3G26770.1
AT3G26760.1
Oeu039798.1
SIN 1004135
SIN 1008273
SIN 1003923
AT3G51680.1
Oeu062066.1
AT1G52340.1
SIN 1017619
Oeu017698.1
Oeu017700.1
Oeu017703.1
SIN 1013074
Oeu002295.1
Oeu009320.2
SIN 1013075
SIN 1013024
Oeu007211.1
SIN 1013073
SIN 1010592
Oeu024218.1
Oeu009071.1
Oeu009058.1
Oeu009065.1
Oeu009063.1
Oeu009055.1
SIN 1022860
AT1G10310.1
SIN 1023601
AT3G04000.1
AT5G18210.1
AT3G03980.1
SIN 1006307
AT4G13180.1
Oeu035806.1
AT1G24360.1
SIN 1026528
Oeu020313.1
Oeu041688.5
Oeu038668.1
SIN 1020075
SIN 1020076
SIN 1018726
SIN 1018725
SIN 1020074
SIN 1020077
Oeu027948.1
AT5G65205.1
AT5G10050.1
SIN 1004351
SIN 1015265
SIN 1024850
AT4G13250.1
SIN 1012305
AT1G24470.1
SIN 1004024
AT1G67730.1
AT5G50690.1
AT5G50590.1
AT5G50700.1
AT5G50600.1
SIN 1022131
SIN 1022130
Oeu048915.2
SIN 1004160
AT3G01980.4
AT3G01980.1
AT3G01980.3
AT3G01980.2
SIN 1026930
AT4G05530.1
Oeu030909.1
Oeu057193.1
Oeu001506.1
SIN 1012223
Oeu006874.1
SIN 1011168
SIN 1023758
AT2G29260.1
Oeu007665.1
SIN 1024217
SIN 1026848
AT3G12800.1
AT3G05260.1
AT1G54870.1
Oeu029689.1
Oeu029688.1
SIN 1025770


Oeu043936.2
SIN 1007158
AT2G04540.1
SIN 1001249
AT5G46290.2
AT5G46290.3
AT5G46290.1
SIN 1019614
SIN 1001803
Oeu042061.1
Oeu007428.1
SIN 1010468
AT1G74960.3
AT1G74960.2
AT1G74960.1
Oeu055336.2
Oeu010830.1
SIN 1024652
SIN 1007554
Oeu000294.1
Oeu008522.1
0.05

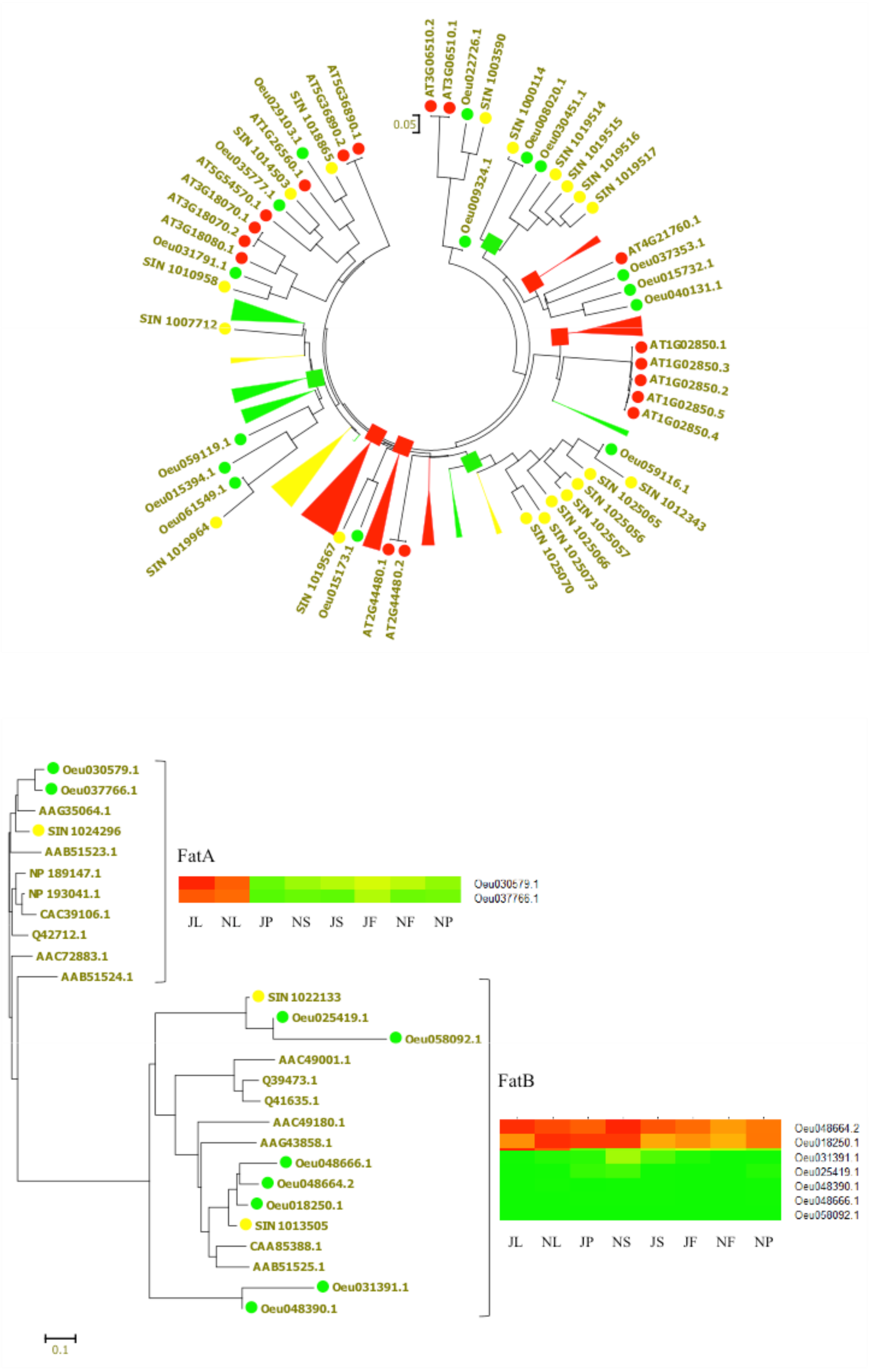
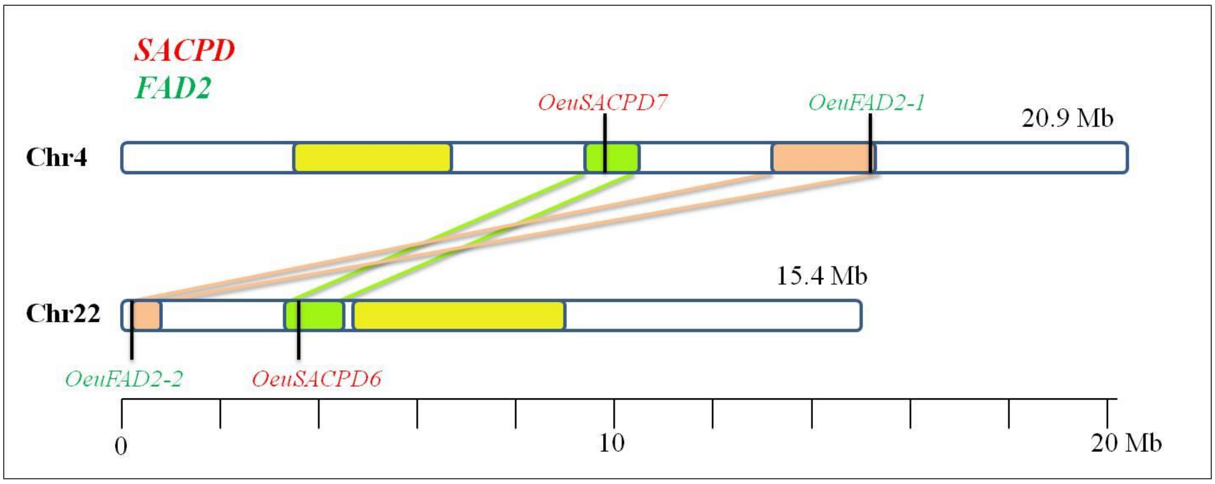
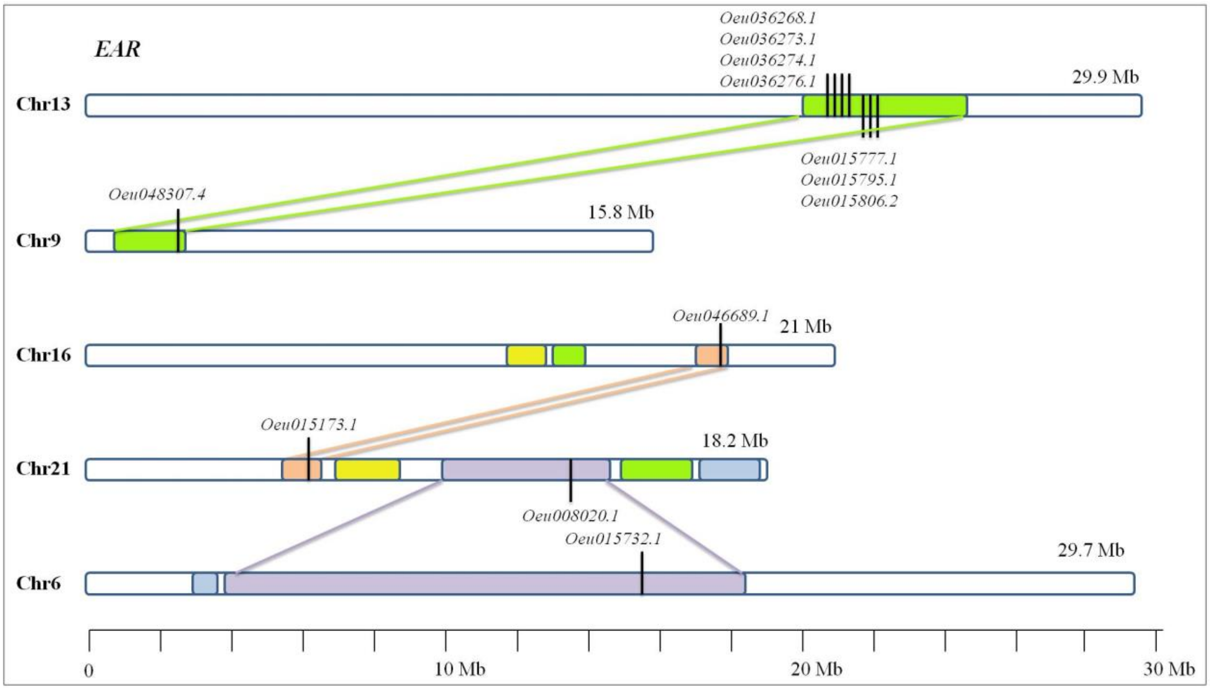
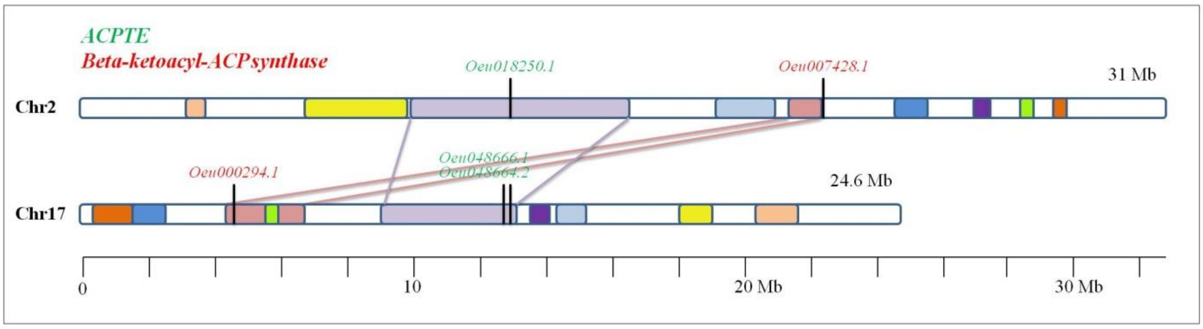


Figure S28. Genome duplication and contribution thereof to gene expansion of the *FabG, KASII, EAR* and *ACPTE* genes in oleaster, sesame and Arabidopsis, colored with green, yellow and red, respectively. **A)** *FabG*, **B)** *KASII,* **C)** *EAR* and **D)** *ACPTE* (*FatA* and *FatB*) genes are coded as circles. Species-specific branches were represented with triangles labeled with the same aforementioned color codes. Squares on the branches represent genome duplication of olive-sesame (green) and Arabidopsis (red).


SACPD
FAD2
OeuSACPD7
OeuFAD2-1
20.9 Mb
Chr4
15.4 Mb
Chr22
OeuFAD2-2
OeuSACPD6
0
10
20 Mb


EAR
Oeu036268.1
Oeu036273.1
Oeu036274.1
Oeu036276.1
29.9 Mb
Chr13
Oeu015777.1
Oeu015795.1
Oeu015806.2
Oeu048307.4
15.8 Mb
Chr9
Oeu046689.1
21 Mb
Chr16
Oeu015173.1
18.2 Mb
Chr21
Oeu008020.1
Oeu015732.1
29.7 Mb
Chr6
0
10 Mb
20 Mb
30 Mb


ACPTE
Beta-ketoacyl-ACPsynthase
Oeu018250.1
Oeu007428.1
31 Mb
Chr2
Oeu048666.1
Oeu048664.2
Oeu000294.1
24.6 Mb
Chr17
0
10
20 Mb
30 Mb

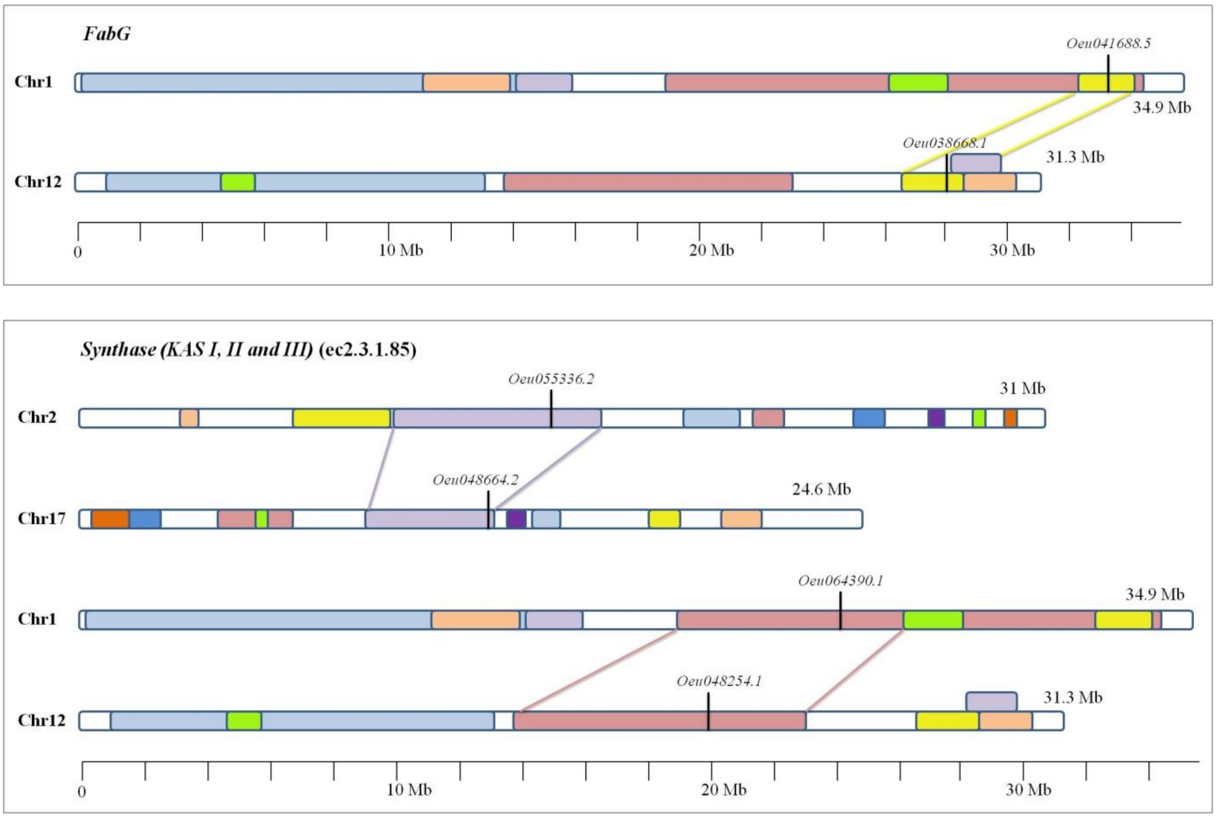
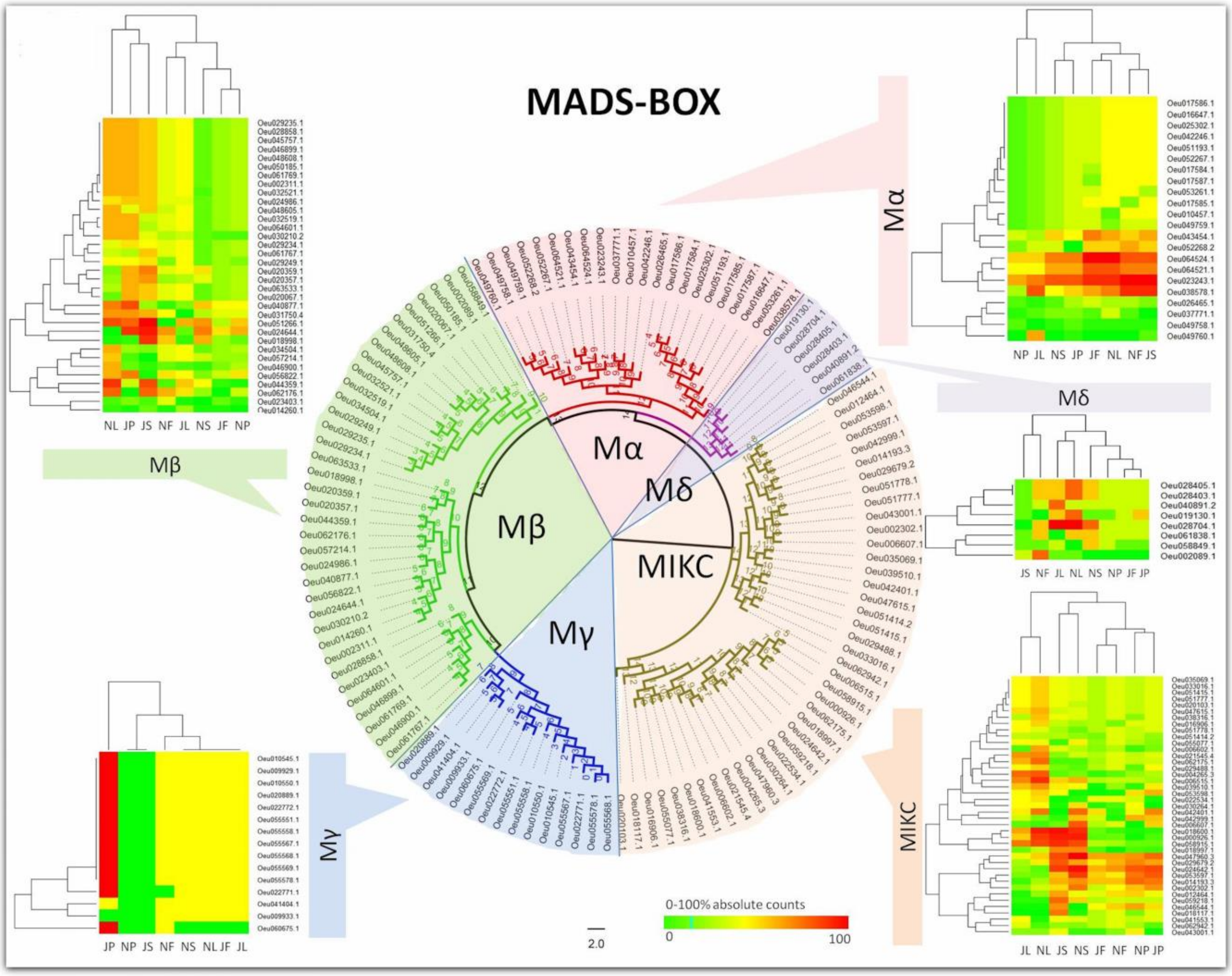


Figure S29. Genome duplication and its contribution to gene expansion of key oil biosynthesis genes in oleaster. Syntenic representation of WGD duplicated (**a**) *FAD2* and *SAPCD* genes located on chromosomes 4 and 22, (**b**) *EAR* genes on chromosomes 13-9, 16-21, and 6-21, (**c**) *ACPTE* and *Beta-ketoacyl ACP-synthase* genes on chromosomes 2 and 17, (**d**) *FabG* genes located on chromosomes 1-17, (**e)** *Synthase (KAS I, II, III)* genes shared by chromosomes 2-17 and 1-12. Color bars represent synteny blocks between oleaster chromosomes.


MADS-BOX
Mα
Mβ
Mγ
Mδ
MIKC
NL JP JS NF JL NS JF NP
Oeu029235.1
Oeu028858.1
Oeu045757.1
Oeu046899.1
Oeu048608.1
Oeu050185.1
Oeu061769.1
Oeu002311.1
Oeu032521.1
Oeu024986.1
Oeu048605.1
Oeu032519.1
Oeu064601.1
Oeu030210.2
Oeu029234.1
Oeu061767.1
Oeu029249.1
Oeu020359.1
Oeu020357.1
Oeu063533.1
Oeu020067.1
Oeu040877.1
Oeu031750.4
Oeu051266.1
Oeu024644.1
Oeu018998.1
Oeu034504.1
Oeu057214.1
Oeu046900.1
Oeu056822.1
Oeu044359.1
Oeu062176.1
Oeu023403.1
Oeu014260.1
NP JL NS JP JF NL NF JS
Oeu017586.1
Oeu016647.1
Oeu025302.1
Oeu042246.1
Oeu051193.1
Oeu052267.1
Oeu017584.1
Oeu017587.1
Oeu053261.1
Oeu017585.1
Oeu010457.1
Oeu049759.1
Oeu043454.1
Oeu052268.2
Oeu064524.1
Oeu064521.1
Oeu023243.1
Oeu038578.1
Oeu026465.1
Oeu037771.1
Oeu049758.1
Oeu049760.1
JS NF JL NL NS NP JF JP
Oeu028405.1
Oeu028403.1
Oeu040891.2
Oeu019130.1
Oeu028704.1
Oeu061838.1
Oeu058849.1
Oeu002089.1
JP NP JS NF NS NL JF JL
Oeu010545.1
Oeu009929.1
Oeu010550.1
Oeu020889.1
Oeu022772.1
Oeu055551.1
Oeu055558.1
Oeu055567.1
Oeu055568.1
Oeu055569.1
Oeu055578.1
Oeu022771.1
Oeu041404.1
Oeu009933.1
Oeu060675.1
JL NL JS NS JF NF NP JP
Oeu035069.1
Oeu033016.1
Oeu051415.1
Oeu051777.1
Oeu020103.1
Oeu047615.1
Oeu038316.1
Oeu016906.1
Oeu051778.1
Oeu051414.2
Oeu055077.1
Oeu006602.1
Oeu021545.4
Oeu062175.1
Oeu029488.1
Oeu004265.3
Oeu006515.1
Oeu039510.1
Oeu053598.1
Oeu022534.1
Oeu030264.1
Oeu042401.1
Oeu042999.1
Oeu006607.1
Oeu018600.1
Oeu000926.1
Oeu058915.1
Oeu018997.1
Oeu047960.3
Oeu029679.2
Oeu024642.1
Oeu053597.1
Oeu014193.3
Oeu002302.1
Oeu012464.1
Oeu059218.1
Oeu046544.1
Oeu018117.1
Oeu041553.1
Oeu062942.1
Oeu043001.1
0-100% absolute counts
0
100
2.0

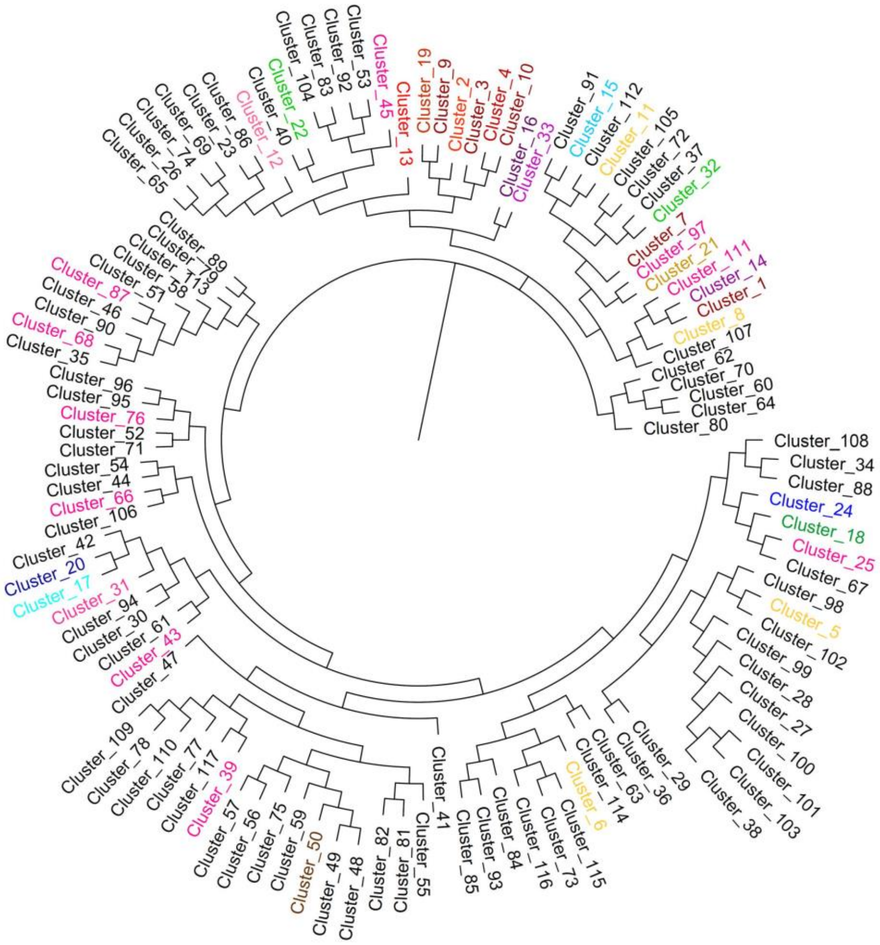


Figure S30. ***MADS-Box*** **gene expansion and expression analyses.** A) Phylogenetic analysis of MADS-Box proteins and their distribution to groups with heatmap analysis. B) Phylogenetic analysis of MADS-Box clusters from 10 selected species using the Neighbor-Joining method (black colored clusters are olive specific). J: July, N: November, L: Leaf, S: Stem, P: Pedicel, F: Fruit.

Tables

Table S1. Summary of library construction and sequencing.

Paired-end libraries	**Insert size**	**Library**	**Read length (bp)**	**Total raw data (Gbp)**	**Total cleaned data (Gbp)**	**Sequence depth (X)***
	250 bp	3	150	106.64	96.83	66.78
	500 bp	2	100	74.95	64.42	44.43
	800 bp	3	100	67.83	57.99	39.99
	2 kbp	3	49	71.65	46.95	32.38
	5 kbp	4	49	64.23	25.80	17.79
	10 kbp	4	49	88.45	18.86	13.01
	20 kbp	4	49	41.92	8.54	5.89
Total	–	23	–	515.67	319.39	220.27

*Sequencing coverage was estimated assuming the genome size of oleaster as 1.46 Gbp by flow cytometry (Fig. S1) and k-mer (Table S2) analyses.

Table S2. 17-mer statistics.

Tables

K	K-mer numeration	Peak depth	Genome size (bp)	Used bases (bp)	Used reads	Coverage X
17	61,190,425,479	42	1,456,914,892	68,921,269,350	459,475,129	47.31

Table S3. Statistics of assembled sequences.

	Contig		Scaffold	
Parameter	**Size (bp)**	**Number**	**Size (bp)**	**Number**
N90	139	1,149,061	144	1,023,435
N80	187	298,764	243	146,631
N70	2,893	41,338	16,388	5,432
N60	12,748	19,000	122,393	2,325
N50	25,485	11,497	228,620	1,448
Longest	694,342	NA	4,441,719	NA
Total size	1,374,783,852	NA	1,485,849,586	NA
Total number (≥100 bp)	NA	2,356,597	NA	2,309,464
Total number (≥1 kbp)	NA	80,929	NA	42,754
Total number (≥2 kbp)	NA	51,199	NA	19,775

NA: not available.

Table S4. Statistics of best match results.

Type	**Total scaffold length (bp)**	**Match length (bp)**	**Match rate (%)**
NT database	136,116,963	14,415,803	10.59
Oleaster	136,116,963	10,811,397	7.94
Repeat	136,116,963	10,726,778	7.88
Oleaster repeat	136,116,963	10,449,855	7.68
rRNA	136,116,963	591,402	0.43
Chloroplast	136,116,963	69,478	0.05
Mitochondria	136,116,963	107,184	0.08

Table S5A. Gene-coverage assessment by RNA-seq data.

Dataset	**Number**	**Total length (bp)**	**Bases covered by assembly (%)**	**Sequences covered by assembly (%)**	**With >90 %**		**With >50 %**	
					Number	%	Number	%
>0 bp	212,714	312,720,638	88.27	91.49	129,227	60.75	182,949	86.01
>200 bp	212,714	312,720,638	88.27	91.49	129,227	60.75	182,949	86.01
>500 bp	145,666	292,175,438	89.02	96.57	82,884	56.90	131,026	89.95
>1000 bp	110,779	266,788,406	89.42	98.43	59,875	54.05	101,185	91.34

Table S5B. Gene-model comparisons among different plant species.

Species	Genome size (Mbp)	Gene	Transcript (CDS)	Transcript length (Mb)	Exon	Intron	TE (Mbp)	TE (%)
Arabidopsis thaliana	119.70	27,407	32,667	38.93	139,382	112,745	23.87	19.90
Eucalyptus grandis	640.10	36,449	36,376	49.92	217,063	171,900	320.6	50.10
Glycine max	955.05	54,257	73,319	132.77	331,060	275,273	587.10	61.47
Oryza sativa	382.80	40,718	97,751	74.14	50,078	190,932	97.99	25.60
Populus trichocarpa	422.90	41,434	45,778	64.48	224,259	179,226	123.71	29.25
Sesamum indicum	273.60	27,148	27,148	86.08	128,461	101,313	77.86	28.46
Solanum tuberosum	844.00	34,998	57,190	79.64	135,708	96,677	452.45	62.20
Vitis vinifera	486.20	26,238	29,971	39.89	156,765	130,419	239.96	49.40
Utricularia gibba	81.87	29,666	29,666	29,18	118,349	90,739	2.50	3.10
Mimulus guttatus	321.72	28,140	28,140	32,92	164,926	137,425	80.43	25.00
Fraxinus excelsior	875.24	38,852	50,743	87,58	93,019	174,657	31.42	35.90
Olea europaea* var. *sylvestris	1,485.85	50,684	50,684	52.70	235,149	184,465	638.29	42.95

Genome data were downloaded from Phytozome 11.0 <https://phytozome.jgi.doe.gov/pz/portal.html> except *S. indicum* from Sinbase <http://ocri-genomics.org/Sinbase>, *F. excelsior* from NCBI (accession no: CBXU010000001) and *U. gibba* from Comparative Genomics (CoGe) <https://genomevolution.org/CoGe/OrganismView.pl?oid=36222>.

Table S6. Repeat-element annotation of oleaster genome (1,485,849,586 bp).

Type	**Repeat Size**	**Genome (%)**
Trf	266,977,162	17.97
RepeatMasker	148,767,573	10.01
ProteinMask	178,304,089	12.00
De novo	696,578,774	46.88
Total	753,310,120	50.70
	De novo	
Type	**Length (bp)**	**Genome (%)**
DNA	56,370,129	3.80
LINE	10,229,171	0.69
SINE	620,866	0.049
LTR	589,401,085	39.67
	Other	
Satellite	4,706,920	0.317
Simple repeat	69,576,976	0.001
Unknown	17,141,899	1.15
Total	696,578,774	46.88

Table S7. Characterization of oleaster TE.

	Repbase TE		TE proteins		*De novo*		Combined TE	
Type	**Length (bp)**	**Genome (%)**	**Length (bp)**	**Genome (%)**	**Length (bp)**	**Genome (%)**	**Length (bp)**	**Genome (%)**
DNA	11,235,341	0.756156	18,332,027	1.233774	56,370,129	3.793798	69,083,581	4.649433
LINE	1,989,299	0.133883	4,658,880	0.313550	10,229,171	0.688439	14,115,865	0.950020
SINE	24,405	0.001642	-	-	620,866	0.041785	643975	0.043341
LTR	136,164,248	9.164067	155,301,526	10.452035	589,401,085	39.667614	599,841,872	40.370296
Other	40,508	0.002726	-	0.000000	-	-	40,508	0.002726
Unknown	-	-	47,387	0.003189	17,141,899	1.153677	171,89,286	1.156866
Total	148,767,573	10.012290	178,304,089	12.000144	622,294,878	41.881418	638,292,541	42.958086

Table S8. General statistics for protein-coding genes.

Gene set		Number	Average transcript length (bp)	Average CDS length (bp)	Average exon per gene	Average exon length (bp)	Average intron length (bp)
De novo	*Augustus*	48,712	2,466.82	1,093.86	4.75	230.28	366.11
	Glimmer HMM	57,696	1,691.59	792.80	3.07	257.93	433.43
Homolog	*S. indicum*	53,389	2,858.96	882.26	3.67	240.52	740.86
	S. tuberosum	63,740	2,064.76	717.90	2.95	243.75	692.39
	V. vinifera	47,834	3,132.16	848.87	3.99	212.87	764.23
	A. thaliana	44,562	2,843.99	888.56	3.86	230.45	684.72
	GLEAN	60,455	2,394.93	973.92	4.17	233.54	448.24
	RNA-seq (Final set*)	50,684	2,767.36	1,040.90	4.63	280,39	470.95

*Genes that can be mapped to known transposable proteins (60% identity) were removed.

Table S9. RNA-seq data statistics.

Sample name	**Clean reads (bp)**	**Size (Gbp)**	**Genome-map rate (%)**	**Expressed genes**
JF	70,829,454	6.37	84.85	49,232
JL	72,219,516	6.50	81.31	46,544
JP	70,141,820	6.31	81.23	51,561
JS	70,962,978	6.39	83.11	41,559
NF	70,369,944	6.33	82.55	49,753
NL	70,481,616	6.34	84.26	51,648
NP	70,440,738	6.34	82.51	51,197
NS	70,857,480	6.38	83.99	52,819

J: July, N: November, F: Fruit, L: Leaf, P: Pedicel and S: Stem.

Table S10. Summary of non-coding RNA in oleaster genome.

Type		Copy number	Average length (bp)	Total length (bp)	Genome (%)
miRNA		441	113.33	49,979	0.003364
tRNA		798	74.83	59,716	0.004019
rRNA	**rRNA**	773	157.71	121,906	0.008204
	18S	130	408.72	53,134	0.003576
	28S	85	103.73	8,817	0.000593
	5.8S	23	118.04	2,715	0.000183
	5S	535	106.99	57,240	0.003852
snoRNA	**H/ACA box**	28	116.96	3,275	0.000220
	Splicing	96	144.15	13,838	0.000931
snRNA		422	113.12	47,737	0.003213
C/D box		298	102.77	30,624	0.002061

Table S11. Small-RNA categorization in oleaster.

	JF	JL1	JL2	NF	NL1	NL2
All reads (≥0, unique)	7,933,475	6,003,166	6,001,443	7,423,620	5,479,750	5,535,758
All reads (≥0, total)	15,260,014	16,950,209	15,931,860	13,817,321	15,153,468	15,710,421
All reads (≥3, unique)	708,454 (8.93%)	767,326 (12.78%)	767,980 (12.80%)	539,834 (7.27%)	715,122 (13.05%)	689,190 (12.45%)
All reads (≥3, total)	7,033,320 (46.09%)	11,013,777 (64.98%)	9,963,199 (62.54%)	6,207,338 (44.92%)	9,668,270 (63.80%)	10,194,673 (64.89%)
Mapped reads (≥3, unique)	563,872 (79.59%)	554,685 (72.9%)	569,266 (74.13%)	437,567 (81.06%)	523,386 (73.19%)	501,089 (72.17%)
Mapped reads (≥3, total)	6,112,550 (86.91%)	8,853,869 (80.39%)	8,179,727 (82.10%)	5,520,095 (88.93%)	7,770,114 (80.37%)	8,329,268 (81.70%)
tRNA (≥3, unique)	1,439 (0.20%)	1,712 (0.22%)	2,186 (0.28%)	1,531 (0.28%)	1,702 (0.24%)	1,989 (0.29%)
tRNA (≥3, total)	34,682 (0.49%)	83,342 (0.76%)	93,483 0.94%)	65,822 (1.06%)	93,672 (0.97%)	144,080 (1.41%)
rRNA (≥3, unique)	4,905 (0.69%)	9,276 (1.21%)	7,510 (0.98%)	5,141 (0.95%)	5,665 (0.79%)	7,740 (1.12%)
rRNA (≥3, total)	204,204 (2.90%)	318,676 (2.89%)	219,929 (2.21%)	207,762 (3.35%)	230,591 (2.39%)	298,633 (2.93%)
snRNA (≥3, unique)	126 (0.02%)	233 (0.03%)	177 (0.02%)	107 (0.02%)	270 (0.04%)	257 (0.04%)
snRNA (≥3, total)	914 (0.01%)	1906 (0.02%)	1,224 (0.01%)	1,025 (0.02%)	2,235 (0.02%)	1,786 (0.02%)
miRNA (≥3, unique)	1,003 (0.14%)	1,695 (0.22%)	1,748 (0.23%)	1,113 (0.21%)	1,364 (0.19%)	1,788 (0.26%)
miRNA (≥3, total)	1,165,170 (16.57%)	1,496,583 (13.59%)	1,069,771 (10.74%)	1,345,870 (21.68%)	1,006,284 (10.41)	1,767,767 (17.34%)

Table S12. Function annotation of miRNA and their targets in oleaster.

Function type	miRNA-target pairs	miRNA	Target	Cellular component	Biological process	Molecular function	Pathw
Development	31	14	12	16	30	10	2
Hormone	52	16	19	5	10	2	0
Metabolism	630	119	269	59	246	172	85
Signal transduction	184	38	52	23	67	18	1
Stress response	1,937	60	236	19	46	40	14
Transcription factor	4,606	111	382	31	203	33	2

Table S13. Functional-annotation statistics.

		Number	Percentage
Total		50,684	–
Annotated	**InterPro**	35,041	69.13
	GO	36,707	72.42
	KEGG	25,416	50.14
	Swiss-Prot	32,418	63.96
Unannotated		13,977	27,58

Table S14a. Analyses of gene families clustered by OrthoMCL.

Species	Genes	Genes in family	Unclustered genes	# Families	Unique families	Unique genes	Average gene number
A. thaliana	27,407	23,388	4,019	13,434	911	3,239	1.74
E. grandis	36,449	28,923	7,526	14,352	974	3,421	2.01
F. excelsior	38,949	30,631	8,318	17,080	325	721	1.80
G. max	54,257	45,116	9,141	15,980	2,077	5,602	2.82
M. guttatus	28,140	24,280	3,860	15,001	585	1,836	1.62
O. sativa	40,718	27,463	13,255	13,582	2,533	9,270	2.02
P. trichocarpa	41,434	33,553	7,881	15,415	1,191	3,519	2.18
S. indicum	27,148	23,572	3,576	14,065	496	3,214	1.68
S. tuberosum	34,998	27,933	7,065	15,035	1,328	4,885	1.86
U. gibba	29,666	21,855	7,811	10,736	716	7,136	2.03
V. vinifera	26,238	19,495	6,743	13,400	672	1,825	1.45
O. europaea* var. *sylvestris	50,684	39,849	10,835	17,208	1,070	8,986	2.32

Table S14b. Orthologous and paralogous gene-model comparisons of 12 plant species clustered by OrthoMCL.

Species	**Number of genes**	**Single copy orthologs**	**Multicopy orthologs**	**Unique paralogs**	**Other orthologs**	**Unclustered**
A. thaliana	27,407	3,514	6,212	3,239	10,423	4,019
E. grandis	36,449	3,882	6,894	3,421	14,726	7,526
F. excelsior	38,949	2,278	11,166	721	16,466	8,318
G. max	54,257	755	17,297	5,602	21,462	9,141
M. guttatus	28,140	3,686	6,007	1,836	12,751	3,860
O. sativa	40,718	3,693	5,730	9,270	8,770	13,255
P. tricocharpa	41,434	2,093	10,937	3,519	17,004	7,881
S. indicum	27,148	3,600	5,904	3,214	10,854	3,576
S. tuberosum	34,998	3,595	6,527	4,885	12,926	7,065
U. gibba	29,666	3,928	4,438	7,136	6,353	7,811
V. vinifera	26,238	4,203	4,270	1,825	9,197	6,743
O. europaea* var. *sylvestris	50,684	2,946	8,774	8,986	19,143	10,835

Table S15. Phylogenetic analyses of duplicates from the older oleaster WGD

	Independent WGD	Shared WGD
# Gene families	188	181
# (Sub)topologies	194	185
# (Sub)topologies consistent with *S. indicum* anchor pair(s)	194	90
# Exclusive ML topologies (AU test, *P* < 0.05)	81	16

Table S16. Synteny relationships identified between chromosomes of *O. europaea* var. *sylvestris* and four other species (*S. indicum, V. vinifera, P. trichocarpa* and *S. tuberosum*). The table displays the number of syntenic blocks and orthologous genes between species.

O. europaea var. *sylvestris*	*S. indicum*		*V. vinifera*		*P. trichocarpa*		*S. tuberosum*	
	#Blocks	**#Genes**	**#Blocks**	**#Genes**	**#Blocks**	**#Genes**	**#Blocks**	**#Genes**
Chr1	36	5,824	31	3,259	65	6,315	35	4,486
Chr2	37	4,047	23	2,673	31	3,793	29	3,271
Chr3	31	4,020	26	2,073	53	6,276	16	3,853
Chr4	40	2,850	23	2,435	47	4,708	30	3,784
Chr5	15	2,101	7	883	22	2,608	13	1,587
Chr6	28	4,364	23	2,865	38	5,497	25	5,161
Chr7	47	4,442	24	3,527	47	5,324	33	3,887
Chr8	14	2,321	17	1,216	27	4,785	15	2,534
Chr9	18	1,938	12	1,201	22	3,418	12	1,187
Chr10	50	7,975	33	5,939	76	9,560	39	9,186
Chr11	62	6,904	46	2,658	62	8,185	50	5,681
Chr12	32	5,017	32	2,662	59	6,651	27	6,461
Chr13	33	3,599	24	1,930	45	4,478	25	3,709
Chr14	24	2,990	15	2,069	32	3,611	13	1,334
Chr15	30	5,570	28	3,959	56	7,288	26	3,894
Chr16	26	5,228	23	2,601	42	7,504	21	4,717
Chr17	32	3,113	18	1,936	31	3,479	21	1,927
Chr18	50	4,626	30	3,835	53	4,621	35	5,221
Chr19	27	1,559	14	942	24	2,122	15	2,360
Chr20	20	1,779	11	1,218	21	1,862	18	2,102
Chr21	11	2,842	8	1,581	22	1,630	11	2,598
Chr22	20	2,027	16	1,325	25	1,298	18	1,989
Chr23	17	1,127	13	828	17	1,187	16	1,418

Table S17. Expansion of *KASI* to *III*, *SACPD*, *EAR*, *FabG* and *FAD2* genes in different plant species.

Species	*KAS I*	*KAS II*	*KAS III*	*KAS I* to *III*	*EAR*	*FabG*	*SACPD*	*FAD2*
O. europaea* var. *sylvestris	6	7	4	17	52	34	7	5
S. indicum	12	7	3	24	41	40	2	2
S. tuberosum	26	11	3	40	32	78	10	16
V. vinifera	7	5	2	14	37	48	11	7
E. grandis	10	10	3	23	52	96	16	6
A. thaliana	9	7	2	18	71	75	8	6
G. max	32	21	3	59	75	92	5	17
P. trichocarpa	27	15	1	43	73	128	10	12
P. persica	13	9	7	29	113	67	5	8
C. canephora	15	13	3	31	57	57	8	9
T. cacao	19	14	2	35	44	65	9	10
Total	176	124	33	333	647	780	91	98

Table S18. Alternate bearing gene annotations in 12 plant species including oleaster.

Gene	*Pfam*	*Arabidopsis thaliana*	*Oryza sativa*	*Prunus persica*	*Populus trichocarpa*	*Vitis vinifera*	*Glycine max*	*Eucalyptus grandis*	*Solanum tuberosum*	*Olea europaea* var. *sylvestris*	*Sesamum indicum*	*Malus domestica*	*Citrus sinensis*
2-Oxoglutarate (2OG)-Fe(II) oxygenase superfamily	PF03171	121	67	119	160	153	244	199	145	141	108	–	–
Phosphatidylethanolamine-binding protein (PEBP; IPR008914)	PF01161	7	19	6	9	4	24	7	12	12	11	17	11
Aminotransferase 1 and 2 (AT1 and AT2) [aminocyclopropane carboxylate (ACC) synthase (ACS) and ACC oxidase (ACO)]	PF00155	41	37	33	55	48	65	40	39	42	38	–	–
Spermine synthase 1 (SPDS1)	PF01564	4	10	6	15	5	31	11	7	7	6	15	12
Polyamine oxidase (PAO; flavin-containing amine oxidoreductase)	PF01593	16	15	14	31	19	32	23	42	24	21	–	–
Diamine oxidase (DAO ; FAD-dependent oxidoreductase)	PF01266	9	8	14	9	7	14	12	10	16	9	–	–
Wall-associated kinase (WAK)	PF08488	19	60	21	55	14	14	36	19	30	14	32	11
Activator protein-1 (AP1)	–	1		1	2	2	4	1	3	3	2	1	1

Table S19. Distribution of transcription factors in oleaster and 10 other plant genomes.

Species TF	***Olea europaea* var. *sylvestris***	***Sesamum indicum***	***Solanum tuberosum***	***Vitis vinifera***	***Eucalyptus grandis***	***Arabidopsis thaliana***	***Glycine max***	***Populus trichocarpa***	***Prunus persica***	***Oryza sativa***	***Theobroma cacao***	**Total**
AP2	**25**	19	45	20	31	18	76	26	19	29	25	333
ARF	**41**	27	20	20	33	22	57	37	17	27	43	344
ARR-B	**32**	18	21	12	9	21	42	63	12	11	18	259
bHLH	**206**	170	206	115	178	225	480	379	133	211	200	2,503
bZIP	**100**	72	95	47	88	127	266	215	50	140	107	1,307
ERF	**214**	105	185	88	136	139	330	209	107	163	107	1,783
FAR1	**105**	22	4	17	46	26	103	111	78	133	92	737
G2-like	**91**	57	85	36	63	40	79	60	36	62	66	675
GATA	**35**	26	50	19	30	41	70	76	22	32	36	437
GeBP	**13**	18	14	1	2	22	9	8	8	13	6	114
GRAS	**82**	53	71	43	95	37	139	151	49	69	70	859
GRF	**15**	11	12	8	7	9	31	26	10	19	17	165
HB-other	**8**	15	12	7	12	11	31	34	7	17	25	179
HB-PHD	**2**	2	3	2	1	3	11	10	2	1	4	41
HD-zip	**51**	38	77	33	52	58	140	114	33	61	56	713
HRT-like	**1**	1	2	1	1	2	1	5	1	1	1	17
HSF	48	0	41	19	37	25	61	47	21	25	30	354
LBD	54	1	47	44	37	50	111	70	43	43	36	536
LFY	**2**	1	1	1	1	1	2	1	1	2	1	14
LSD	**11**	5	12	4	3	13	18	18	5	11	8	108
MADS I	**72**	72	137	54	97	108	172	102	78	69	69	1,066
MADS II	**48**	31	35	39	51	41	93	58	32	39	34	490

MYB	**150**	125	126	138	174	168	369	266	121	130	138	1,905
MYB-related	**263**	224	232	243	268	258	601	374	215	235	226	3,139
NAC	**187**	128	107	71	164	113	180	170	115	140	107	1,482
NF-X1	**1**	1	1	1	3	1	4	2	1	2	2	19
NF-YA	**17**	8	11	7	8	10	21	13	6	11	7	119
NF-YB	**38**	28	32	22	24	27	66	39	22	29	23	350
NF-YC	**16**	21	12	8	12	21	42	27	9	19	31	218
Nin-like	**68**	38	54	30	39	52	100	80	40	43	41	585
NZZ/SPL	**2**	2	0	1	0	1	0	4	2	0	2	14
RAV	**236**	210	198	101	156	145	125	210	124	159	123	1,787
S1Fa-like	**10**	2	7	2	1	4	4	2	1	2	1	36
SAP	**3**	3	1	1	1	1	2	1	1	0	1	15
SBP	**35**	29	31	19	25	30	73	68	17	29	26	382
SRS	**8**	8	7	5	7	16	33	21	6	6	8	125
STAT	**9**	5	1	1	0	4	1	4	1	1	5	32
TALE	**24**	21	32	21	29	33	101	80	22	45	29	437
TCP	**52**	33	43	15	16	33	71	60	19	23	31	396
Trihelix	**45**	37	39	26	25	34	93	78	33	40	40	490
VOZ	**2**	2	4	2	3	3	20	8	3	2	8	57
Whirly	**3**	2	4	2	9	4	13	6	2	2	5	52
WOX	**38**	33	8	11	10	18	42	26	10	17	16	229
WRKY	**135**	70	84	59	79	73	176	102	58	128	61	1,025
Total	**2,496**	**1,793**	**2,121**	**1,353**	**1,989**	**2,013**	**4,287**	**3,344**	**1,528**	**2,173**	**1,916**	**25,038**

Table S20. Disease-resistance genes of oleaster in comparison with 12 other plant species.

			Woody plants							Herbaceous plants					
	Type	**Code**	**Oleaster**	**Cacao**	**Common grape-vine**	**Cassava**	**Poplar**	**Chestnut**	**Eucalyptus**	**Cucumber**	**Potato**	**Arabidopsis**	**Soybean**	**Tomato**	**Sesame**
TIR genes	**TIR-NBS**	TN	**0**	4	3	**0**	10	5	276	2	12	17	n/a*	8	0
	TIR-NBS-LRR	TNL	**0**	8	17	25	78	22	174	11	37	79	116	16	0
	Subtotal 1		**0**	**12**	**20**	5	**88**	**27**	**450**	**13**	**49**	**96**	**116**	**24**	**0**
non-TIR genes	**CC-NBS**	CN	**86**	46	18	23	14	96	107	1	24	8	n/a*	18	25
	CC-NBS-LRR	CNL	**10**	82	28	118	120	32	133	17	65	17	20	4	5
	NBS-LRR	NL	**18**	104	121	**171**	132	320	250	23	177	20	32	21	23
	NBS	N	**176**	53	129	**171**	62	44	277	1	104	26	n/a*	188	118
	Subtotal 2		**290**	**285**	**296**	1	**328**	**492**	**767**	**42**	**370**	**71**	**52**	**231**	**171**
	Grand total		**290**	**297**	**316**	**249**	**416**	**546**	**1,667**	**55**	**468**	**167**	**314**	**255**	**171**
	Reference		**This study**	Yu et al. 2014[2]	Wang et al. 2014[1]	Lozano et al. 2015[3]	Yang et al. 2008[4]	Zhong et al. 2015[5]	Christie et al. 2015[6]	Wan et al. 2013[7]	Lozano et al. 2012[8]	Meyer et al 2003[9]	Kang et al. 2012[10]	Wang et al. 2014[1]	Wang et al. 2014[1]

References

1. 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 Molecular Biology* 2012:6.
2. Dolezel J & Bartos J (2005) Plant DNA flow cytometry and estimation of nuclear genome size. *Ann Bot* 95(1):99-110.
3. Marçais G & Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27(6):764-770.
4. Cruz F*, et al.* (2016) Genome sequence of the olive tree, Olea europaea. *Gigascience* 5:29.
5. Li R*, et al.* (2010) The sequence and de novo assembly of the giant panda genome. *Nature* 463(7279):311-317.
6. Li R*, et al.* (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome research* 20(2):265-272.
7. You MS*, et al.* (2013) A heterozygous moth genome provides insights into herbivory and detoxification. *Nat Genet* 45(2):220-225.
8. Boetzer M, Henkel CV, Jansen HJ, Butler D, & Pirovano W (2011) Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27(4):578-579.
9. Wei W*, et al.* (2011) Characterization of the sesame (Sesamum indicum L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. *BMC genomics* 12:451.
10. Elshire RJ*, et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6(5):e19379.
11. Raman H*, et al.* (2014) Genome-wide delineation of natural variation for pod shatter resistance in Brassica napus. *PLoS One* 9(7):e101673.
12. Voorrips RE (2002) MapChart: software for the graphical presentation of linkage maps and QTLs. *J Hered* 93(1):77-78.
13. Risterucci A*, et al.* (2000) A high-density linkage map of Theobroma cacao L. *Theoretical and*

Applied Genetics 101(5-6):948-955.

14. Pugh T*, et al.* (2004) A new cacao linkage map based on codominant markers: development and integration of 201 new microsatellite markers. *Theoretical and Applied Genetics* 108(6):1151-1161.

15. Fouet O*, et al.* (2011) Structural characterization and mapping of functional EST-SSR markers in Theobroma cacao. *Tree Genetics & Genomes* 7(4):799-817.

16. Li H & Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26(5):589-595.

17. Tang H*, et al.* (2015) ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol* 16:3.

18. Edgar RC & Myers EW (2005) PILER: identification and classification of genomic repeats. *Bioinformatics* 21 Suppl 1:i152-158.

19. McCarthy EM & McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* 19(3):362-367.

20. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* 27(2):573.

21. Tarailo‐Graovac M & Chen N (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*:4.10. 11-14.10. 14.

22. 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.

23. Zhong S*, et al.* (2011) High-throughput illumina strand-specific RNA sequencing library preparation. *Cold Spring Harb Protoc* 2011(8):940-949.

24. Grabherr MG*, et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* 29(7):644-652.

25. Kim D, Langmead B, & Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 12(4):357-360.

26. Quinlan AR & Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features.

Bioinformatics 26(6):841-842.

27. Mortazavi A, Williams BA, McCue K, Schaeffer L, & Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5(7):621-628.

28. Ashburner M*, et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat Genet* 25(1):25-29.

29. Elsik CG*, et al.* (2007) Creating a honey bee consensus gene set. *Genome Biology* 8(1).

30. 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.

31. Birney E, Clamp M, & Durbin R (2004) GeneWise and genomewise. *Genome research* 14(5):988-995.

32. Majoros WH, Pertea M, & Salzberg SL (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20(16):2878-2879.

33. Stanke M*, et al.* (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research* 34(suppl 2):W435-W439.

34. Lowe TM & Chan PP (2016) tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res* 44(W1):W54-57.

35. Nawrocki EP & Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29(22):2933-2935.

36. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, & Eddy SR (2003) Rfam: an RNA family database. *Nucleic Acids Research* 31(1):439-441.

37. Yanik H*, et al.* (2013) Genome-wide identification of alternate bearing-associated microRNAs (miRNAs) in olive (Olea europaea L.). *BMC Plant Biol* 13:10.

38. Xie F, Xiao P, Chen D, Xu L, & Zhang B (2012) miRDeepFinder: a miRNA analysis tool for deep sequencing of plant small RNAs. *Plant molecular biology*.

39. Du JL*, et al.* (2014) KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model. *Mol Biosyst* 10(9):2441-2447.

40. Griffiths-Jones S (2004) The microRNA Registry. *Nucleic Acids Res* 32(Database issue):D109-111.

41. Smith TF & Waterman MS (1981) Identification of common molecular subsequences. *J Mol Biol* 147(1):195-197.

42. Xie F, Wang Q, Sun R, & Zhang B (2014) Deep sequencing reveals important roles of microRNAs in response to drought and salinity stress in cotton. *Journal of experimental botany*.

43. Dai X & Zhao PX (2011) psRNATarget: a plant small RNA target analysis server. *Nucleic Acids Res* 39(Web Server issue):W155-159.

44. Boeckmann B*, et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 31(1):365-370.

45. Jones P*, et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30(9):1236-1240.

46. Conesa A*, et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21(18):3674-3676.

47. Ye J*, et al.* (2006) WEGO: a web tool for plotting GO annotations. *Nucleic acids research* 34(suppl 2):W293-W297.

48. Goodstein DM*, et al.* (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* 40(Database issue):D1178-1186.

49. Sollars ES*, et al.* (2017) Genome sequence and genetic diversity of European ash trees. *Nature* 541(7636):212-216.

50. Futami R, Munoz-Pomer, A., Viu,J.M., Dominguez-Escriba,L., Covelli,L., Bernet,G.P., Sempere,J.M., Moya,A. and Llorens,C. (2011) GPRO The professional tool for annotation, management and functional analysis of omic databases. *Biotechvana Bioinformatics* 2011-soft3.

51. Nakaya A*, et al.* (2013) KEGG OC: a large-scale automatic construction of taxonomy-based ortholog clusters. *Nucleic Acids Res* 41(Database issue):D353-357.

52. Li L, Stoeckert CJ, & Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research* 13(9):2178-2189.

53. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* 32(5):1792-1797.

54. Ronquist F*, et al.* (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology* 61(3):539-542.

55. THE ANGIOSPERM PHYLOGENY GROUP, An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Botanical Journal of the Linnean Society, 2016, 181, 1–2056.

56. Tank DC*, et al.* (2015) Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytol* 207(2):454-467.

57. Soltis DE*, et al.* (2011) Angiosperm phylogeny: 17 genes, 640 taxa. *Am J Bot* 98(4):704-730.

58. Ibarra-Laclette E*, et al.* (2013) Architecture and evolution of a minute plant genome. *Nature* 498(7452):94-98.

59. He Y*, et al.* (2016) The Complete Chloroplast Genome Sequences of the Medicinal Plant Pogostemon cablin. *Int J Mol Sci* 17(6).

60. Sanderson MJ (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19(2):301-302.

61. Yang ZH (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24(8):1586-1591.

62. Vanneste K, Van de Peer Y, & Maere S (2013) Inference of genome duplications from age distributions revisited. *Mol Biol Evol* 30(1):177-190.

63. Enright AJ, Van Dongen S, & Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30(7):1575-1584.

64. Goldman N & Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* 11(5):725-736.

65. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24(8):1586-1591.

66. Guindon S*, et al.* (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59(3):307-321.

67. Proost S*, et al.* (2012) i-ADHoRe 3.0--fast and sensitive detection of genomic homology in

extremely large data sets. *Nucleic Acids Res* 40(2):e11.

68. Fostier J*, et al.* (2011) A greedy, graph-based algorithm for the alignment of multiple homologous gene lists. *Bioinformatics* 27(6):749-756.

69. 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.

70. Ostlund G*, et al.* (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res* 38(Database issue):D196-203.

71. Drummond AJ, Suchard MA, Xie D, & Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* 29(8):1969-1973.

72. Heled J & Drummond AJ (2012) Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Syst Biol* 61(1):138-149.

73. Emms DM & Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* 16:157.

74. Capella-Gutierrez S, Silla-Martinez JM, & Gabaldon T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15):1972-1973.

75. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312-1313.

76. Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. *Syst Biol* 51(3):492-508.

77. Shimodaira H & Hasegawa M (2001) CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17(12):1246-1247.

78. Tang H*, et al.* (2008) Synteny and collinearity in plant genomes. *Science* 320(5875):486-488.

79. Soderlund C, Nelson W, Shoemaker A, & Paterson A (2006) SyMAP: A system for discovering and viewing syntenic regions of FPC maps. *Genome Res* 16(9):1159-1168.

80. 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.

81. Kurtz S*, et al.* (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5(2):R12.

82. Krzywinski M*, et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Res* 19(9):1639-1645.

83. Xu Z & Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* 35(Web Server issue):W265-268.

84. Huang X & Madan A (1999) CAP3: A DNA sequence assembly program. *Genome Res* 9(9):868-877.

85. Jurka J*, et al.* (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110(1-4):462-467.

86. Tamura K, Stecher G, Peterson D, Filipski A, & Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* 30(12):2725-2729.

87. Ma J & Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci U S A* 101(34):12404-12410.

88. 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.

89. Altschul SF*, et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(17):3389-3402.

90. Pruitt KD, Tatusova T, & Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35(Database issue):D61-65.

91. Futami R*, et al.* (2011) GPRO The professional tool for annotation, management and functional analysis of omic databases. *Biotechvana Bioinformatics: 2011-SOFT3*.

92. Wang L*, et al.* (2014) Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis. *Genome Biol* 15(2):R39.

93. Conesa A & Gotz S (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int J Plant Genomics* 2008:619832.

94. Ascensión Rueda IS, Manuel Olalla, Rafael Giménez, Luis Lara, and Carmen Cabrera-Vique (2014)

Characterization of Fatty Acid Profile of Argan Oil and Other Edible Vegetable Oils by Gas Chromatography and Discriminant Analysis. *Journal of Chemistry* 2014(843908):1-8.

95. Harwood JL & Guschina IA (2013) Regulation of lipid synthesis in oil crops. *FEBS Lett* 587(13):2079-2081.

96. Bates PD, Stymne S, & Ohlrogge J (2013) Biochemical pathways in seed oil synthesis. *Curr Opin Plant Biol* 16(3):358-364.

97. Posada D (2008) jModelTest: Phylogenetic model averaging. *Molecular Biology and Evolution* 25(7):1253-1256.

98. Akaike H (1981) Citation Classic - a New Look at the Statistical-Model Identification. *Cc/Eng Tech Appl Sci* (51):22-22.

99. Le SQ & Gascuel O (2008) An improved general amino acid replacement matrix. *Mol Biol Evol* 25(7):1307-1320.

100. Felsenstein J (1985) Confidence-Limits on Phylogenies - an Approach Using the Bootstrap. *Evolution* 39(4):783-791.

101. Lior Rokach OM (2005) Data Mining and Knowledge Discovery Handbook. *Data Mining and Knowledge Discovery Handbook*, ed Oded Maimon LR), pp 321-352.

102. Nakaya A*, et al.* (2013) KEGG OC: a large-scale automatic construction of taxonomy-based ortholog clusters. *Nucleic Acids Res.* 41(D1):D353-D357.

103. Vanhercke T, Wood CC, Stymne S, Singh SP, & Green AG (2013) Metabolic engineering of plant oils and waxes for use as industrial feedstocks. *Plant Biotechnol J* 11(2):197-210.

104. Aslan S, Hofvander P, Dutta P, Sitbon F, & Sun C (2015) Transient silencing of the KASII genes is feasible in Nicotiana benthamiana for metabolic engineering of wax ester composition. *Sci Rep* 5:11213.

105. Watterson GA (1983) On the time for gene silencing at duplicate Loci. *Genetics* 105(3):745-766.

106. Lynch M & Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290(5494):1151-1155.

107. Pasquier J*, et al.* (2016) Gene evolution and gene expression after whole genome duplication in

fish: the PhyloFish database. *BMC genomics* 17:368.

108. Salas JJ & Ohlrogge JB (2002) Characterization of substrate specificity of plant FatA and FatB acyl-ACP thioesterases. *Arch Biochem Biophys* 403(1):25-34.

109. Lang D*, et al.* (2010) Genome-Wide Phylogenetic Comparative Analysis of Plant Transcriptional Regulation: A Timeline of Loss, Gain, Expansion, and Correlation with Complexity. *Genome Biol Evol* 2:488-503.

110. Finn RD*, et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research* 44(D1):D279-D285.

111. Wang L, Yu J, Li D, & Zhang X (2015) Sinbase: an integrated database to study genomics, genetics and comparative genomics in Sesamum indicum. *Plant Cell Physiol* 56(1):e2.

112. Wei X*, et al.* (2015) Genome-wide identification and analysis of the MADS-box gene family in sesame. *Gene* 569(1):66-76.

113. Katoh K, Misawa K, Kuma K, & Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30(14):3059-3066.

114. Saitou N & Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4.

115. Klebanov RD & Chuiko MP (1987) [A simple method of evaluating differences in indices of number of days of temporary disability]. *Gig Tr Prof Zabol* (7):39-42.

116. Vrebalov J*, et al.* (2002) A MADS-box gene necessary for fruit ripening at the tomato ripening-inhibitor (rin) locus. *Science* 296(5566):343-346.

117. Irfan M*, et al.* (2016) Fruit Ripening Regulation of alpha-Mannosidase Expression by the MADS Box Transcription Factor RIPENING INHIBITOR and Ethylene. *Front Plant Sci* 7:10.

118. Seymour GB*, et al.* (2011) A SEPALLATA gene is involved in the development and ripening of strawberry (Fragaria x ananassa Duch.) fruit, a non-climacteric tissue. *Journal of experimental botany* 62(3):1179-1188.

119. Schaffer RJ, Ireland HS, Ross JJ, Ling TJ, & David KM (2013) SEPALLATA1/2-suppressed mature apples have low ethylene, high auxin and reduced transcription of ripening-related genes. *AoB*

Plants 5:pls047.

120. Elitzur T*, et al.* (2016) Banana MaMADS Transcription Factors Are Necessary for Fruit Ripening and Molecular Tools to Promote Shelf-Life and Food Security. *Plant Physiol* 171(1):380-391.

121. Dong T*, et al.* (2013) A tomato MADS-box transcription factor, SlMADS1, acts as a negative regulator of fruit ripening. *Plant Physiol* 163(2):1026-1036.

122. Gapper NE, McQuinn RP, & Giovannoni JJ (2013) Molecular and genetic regulation of fruit ripening. *Plant molecular biology* 82(6):575-591.

123. Alagna F*, et al.* (2012) Olive phenolic compounds: metabolic and transcriptional profiling during fruit development. *BMC plant biology* 12(1):1-19.

124. Omar SH (2010) Oleuropein in olive and its pharmacological effects. *Sci Pharm* 78.

125. Kumar S, Stecher G, & Tamura K (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*.

126. Jones DT, Taylor WR, & Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8(3):275-282.

127. Letunic I & Bork P (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research* 44(W1):W242-W245.

128. Caraux G & Pinloche S (2005) PermutMatrix: a graphical environment to arrange gene expression profiles in optimal linear order. *Bioinformatics* 21(7):1280-1281.

129. Gruszka D (2013) The Brassinosteroid Signaling Pathway—New Key Players and Interconnections with Other Signaling Networks Crucial for Plant Development and Stress Tolerance. *International Journal of Molecular Sciences* 14(5):8740-8774.

130. Maeda H & Dudareva N (2012) The Shikimate Pathway and Aromatic Amino Acid Biosynthesis in Plants. *Annual Review of Plant Biology* 63(1):73-105.

131. Ruiz-Sola MÁ & Rodríguez-Concepción M (2012) Carotenoid Biosynthesis in Arabidopsis: A Colorful Pathway. *The Arabidopsis Book / American Society of Plant Biologists* 10:e0158.

132. Turktas M*, et al.* (2013) Nutrition metabolism plays an important role in the alternate bearing of the olive tree (Olea europaea L.). *PLoS One* 8(3):e59876.

133. Dundar E, Sonmez GD, & Unver T (2015) Isolation, molecular characterization and functional analysis of OeMT2, an olive metallothionein with a bioremediation potential. *Mol Genet Genomics* 290(1):187-199.

134. Dundar E, Suakar O, Unver T, & Dagdelen A (2013) Isolation and expression analysis of cDNAs that are associated with alternate bearing in Olea europaea L. cv. Ayvalik. *BMC Genomics* 14:219.

135. Jin JP, Zhang H, Kong L, Gao G, & Luo JC (2014) PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Research* 42(D1):D1182-D1187.

136. Kim MJ, Kim JK, Shin JS, & Suh MC (2007) The SebHLH transcription factor mediates trans-activation of the SeFAD2 gene promoter through binding to E- and G-box elements. *Plant molecular biology* 64(4):453-466.

137. Song QX*, et al.* (2013) Soybean GmbZIP123 gene enhances lipid content in the seeds of transgenic Arabidopsis plants. *Journal of experimental botany* 64(14):4329-4341.

138. Wang HW*, et al.* (2007) The soybean Dof-type transcription factor genes, GmDof4 and GmDof11, enhance lipid content in the seeds of transgenic Arabidopsis plants. *Plant J* 52(4):716-729.

139. Kang NK*, et al.* (2015) Effects of overexpression of a bHLH transcription factor on biomass and lipid production in Nannochloropsis salina. *Biotechnol Biofuels* 8:200.

140. Ibanez-Salazar A*, et al.* (2014) Over-expression of Dof-type transcription factor increases lipid production in Chlamydomonas reinhardtii. *J Biotechnol* 184:27-38.

141. Zhang J*, et al.* (2014) Overexpression of the soybean transcription factor GmDof4 significantly enhances the lipid content of Chlorella ellipsoidea. *Biotechnol Biofuels* 7(1):128.

142. Fu L, Niu B, Zhu Z, Wu S, & Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28(23):3150-3152.

143. Hunter S*, et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* 40(Database issue):D306-D312.

144. Finn RD, Clements J, & Eddy SR (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 39(Web Server issue):W29-37.

145. Delorenzi M & Speed T (2002) An HMM model for coiled-coil domains and a comparison with

PSSM-based predictions. *Bioinformatics* 18(4):617-625.

146. McDonnell AV, Jiang T, Keating AE, & Berger B (2006) Paircoil2: improved prediction of coiled coils from sequence. *Bioinformatics* 22(3):356-358.

147. Lozano R, Ponce O, Ramirez M, Mostajo N, & Orjeda G (2012) Genome-wide identification and mapping of NBS-encoding resistance genes in Solanum tuberosum group phureja. *PLoS One* 7(4):e34775.

148. Tian Y, Fan L, Thurau T, Jung C, & Cai D (2004) The absence of TIR-type resistance gene analogues in the sugar beet (Beta vulgaris L.) genome. *J Mol Evol* 58(1):40-53.

149. Tarr DE & Alexander HM (2009) TIR-NBS-LRR genes are rare in monocots: evidence from diverse monocot orders. *BMC Res Notes* 2:197.